ID N. AIDR01/16880





UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITÀ DEGLI STUDI DI CATANIA

DEPARTMENT OF ELECTRICAL, ELECTRONIC AND COMPUTER ENGINEERING

Italian National Ph.D. in Artificial Intelligence

Health and Life Sciences

XXXVII Cycle

Multimodal Learning to Enhance Diagnosis, Monitoring and Accessibility in Healthcare

Supervisors

Prof. Simone Palazzo

Prof. Danilo Avola

Prof. Concetto Spampinato

Prof. Francesco Rundo

Candidate

Raffaele Mineo

January, 2026

Abstract

Artificial intelligence (AI) is increasingly adopted in diagnosis, monitoring, and assistive technologies, yet real-world impact is often limited by practical constraints on data collection: heterogeneous acquisition protocols, privacy-sensitive modalities, and deployment requirements in resource-constrained settings. This dissertation studies multimodal learning across three independent research tracks in healthcare, showing how combining complementary inputs and modern representation learning can improve robustness, generalization, and usability in realistic workflows.

Part I focuses on cardiovascular analysis from coronary angiography. We develop spatio-temporal deep models that operate on angiographic sequences to quantify stenosis severity and to estimate functional indices (FFR/iFR) without invasive probes. The contributions span 3D and attention-based modeling, hybrid convolutional-transformer designs for temporal contrast-flow representation, and learning under distributed multi-center settings via federated learning, including detection-transformer-based pipelines on key frames. We also investigate how foundation-model paradigms can be adapted to angiography with parameter-efficient strategies, leveraging contrastive vision-language representations and learnable prompts to improve transfer under domain shift.

Part II targets automated cognitive and behavioral assessment from scalable, non-invasive signals compatible with commodity devices and consumer wearables. We study the extraction of subtle facial dynamics through phase-based motion amplification and their integration with affective and task-performance descriptors for estimating cognitive screening scores. We also investigate transformer-based modeling of eye-gaze traces collected during structured tasks, emphasizing temporal representation learning and evaluation protocols aligned with remote assessment. In addition, we include a long-term wearable sensing study that analyzes multimodal smartwatch time series under realistic conditions of missingness and heterogeneity, covering both subject identification and personalized relapse detection in the psychotic spectrum.

Part III addresses accessibility in clinical communication through sign language recognition for medical vocabulary. We consolidate multimodal dataset design and benchmarking for

Italian Sign Language, and we advance privacy-preserving recognition using mmWave radar representations and sequence models. The work also explores semantic grounding through text-aligned embeddings and reports benchmarking outcomes through community evaluation initiatives.

Overall, the dissertation shows that multimodal learning, by aligning and fusing complementary signals and by enabling modality substitution when constraints arise, provides a practical route to robust and deployable healthcare AI across diagnosis, monitoring, and accessibility.

Lay Summary

Artificial intelligence (AI) can support clinicians in diagnosing disease, monitoring patients over time, and improving access to care. In practice, however, many AI solutions remain hard to translate beyond research settings. A common reason is not the lack of advanced algorithms, but the difficulty of obtaining suitable data: some measurements are expensive or invasive, others are available only in specialized centers, and many practical signals raise privacy and deployment concerns.

This dissertation investigates multimodal AI across three parallel research lines in healthcare, each centered on a different clinical need and data source. The common goal is to design methods that remain effective under real-world constraints, including limited availability of specialized sensing, variability across sites and devices, and privacy-sensitive acquisition.

In the first part, the thesis focuses on cardiovascular diagnostics from coronary angiography, a widely used X-ray procedure to visualize coronary vessels. We develop AI models that analyze angiography videos to estimate stenosis severity and to approximate functional indices of lesion significance (FFR/iFR), aiming to support decision-making in intermediate cases. We also consider how such systems can be developed and evaluated across multiple hospitals without centralizing patient data, using privacy-preserving collaborative training.

In the second part, the thesis addresses cognitive and behavioral monitoring using scalable, non-invasive signals acquired with commodity cameras and consumer wearables. We study subtle facial dynamics and eye-gaze patterns recorded during structured tasks to automatically estimate cognitive screening scores, and we complement these contributions with long-term multimodal smartwatch analytics for subject identification and personalized relapse detection in the psychotic spectrum. The emphasis is on approaches that support repeated assessment in everyday conditions, while remaining robust to subject variability, missing data, and heterogeneous acquisition.

In the third part, the thesis tackles accessibility in medical communication for Deaf and hard-of-hearing patients through automatic recognition of sign language in clinical contexts. We introduce and analyze multimodal datasets for Italian Sign Language, and we investigate privacy-preserving recognition using millimeter-wave radar, which can capture

motion patterns without storing identifiable video. We further study how radar signals can be aligned with text labels to support more reliable recognition, and we report results from community benchmarking activities.

Overall, the dissertation shows that progress in clinical AI is not only driven by stronger models, but also by careful choices about what signals to acquire and how to use them responsibly. By combining multimodal learning with privacy-aware design and deployment-oriented evaluation, the thesis aims to make AI tools more practical, reliable, and easier to adopt in real healthcare settings.

Preface

This dissertation collects research conducted during my Ph.D. and is organized around three parallel lines of work in healthcare AI: cardiovascular diagnostics from coronary angiography, cognitive and behavioral monitoring from scalable non-invasive signals acquired with commodity devices and consumer wearables, and sign-language recognition for medical communication with an emphasis on privacy-preserving sensing. Each line is documented through peer-reviewed publications and complementary material that contextualizes the methodological choices and the experimental evidence.

Parts of this thesis build upon work previously published in conference proceedings and journals. In all cases, the text has been adapted and restructured to ensure coherence within a single dissertation narrative, while preserving the core technical content and the scientific contributions of the original papers. Figures and tables are reproduced or redrawn from the corresponding publications when appropriate. Bibliographic references are provided in the thesis reference list.

A portion of the work summarized in this dissertation was also presented and discussed within the MICCAI 2025 Doctoral Consortium [237], where my doctoral project was accepted for the publication of an official report.

The research described here was conducted in collaboration with multiple academic and clinical partners. My contribution spans the conception of methods and experimental protocols, model development and implementation, dataset design and curation, analysis and interpretation of results, and manuscript preparation. When a contribution resulted from joint work, the roles of collaborators are acknowledged in the corresponding sections and in the dedicated acknowledgements.

Some results presented in this dissertation are associated with intellectual property activities. Specifically, patents related to angiography-based estimation of functional indices (FFR/iFR) and to a physiological-monitoring safety device are cited where relevant. For these components, the thesis describes the ideas at an architectural and scientific level, focusing on reproducible principles and validated mechanisms, while avoiding disclosure of implementation details that are not publicly available.

Unless otherwise stated, the experiments reported in this dissertation were performed by the author, and all conclusions reflect the author's interpretation of the results based on the evidence available at the time of writing.

List of Scientific Contributions

This dissertation is based on selected peer-reviewed publications and patent applications authored during the Ph.D. period. Additional scholarly outputs produced during the same period are reported separately.

Peer-reviewed publications included in the dissertation

1. Raffaele Mineo, Federica Proietto Salanitri, Giovanni Bellitto, Ovidio De Filippo, Fabrizio D'Ascenzo, Simone Palazzo, and Concetto Spampinato. Coronary artery stenosis assessment in x-ray angiography through spatio-temporal attention for non-invasive ffr and ifr estimation. In *BIOSTEC (1)*, pages 305–312, 2024 [245]

2. (*) Ovidio De Filippo, Raffaele Mineo, Michele Millesimo, Wojciech Wańha, Federica Proietto Salanitri, Antonio Greco, Antonio Maria Leone, Luca Franchin, Simone Palazzo, Giorgio Quadri, et al. Non-invasive physiological assessment of intermediate coronary stenoses from plain angiography through artificial intelligence: the starflow system. *European Heart Journal-Quality of Care and Clinical Outcomes*, 11(3):343–352, 2025 [79]

3. Raffaele Mineo, F Proietto Salanitri, G Bellitto, I Kavasidis, O De Filippo, M Millesimo, GM De Ferrari, M Aldinucci, D Giordano, S Palazzo, et al. A convolutional-transformer model for ffr and ifr assessment from coronary angiography. *IEEE Transactions on Medical Imaging*, 43(8):2866–2877, 2024 [244]

4. Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Concetto Spampinato, and Simone Palazzo. Learning joint text and visual tokens in clip for medical image analysis. In *Proceedings of the 2nd International Workshop on Multimedia Computing for Health and Medicine*, pages 45–51, 2025 [238]

*The candidate contributed equally and should be considered first author.

5. Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Ulas Bagci, Daniela Giordano, Concetto Spampinato, and Simone Palazzo. Conditional visuo-textual prompt learning for medical image analysis. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025 [239]

6. Raffaele Mineo, Amelia Sorrenti, and Federica Proietto Salanitri. Fedetr: A federated approach for stenosis detection in coronary angiography. In *International Conference on Image Analysis and Processing*, pages 189–200. Springer, 2023 [251]

7. Raffaele Mineo, Federica Proietto Salanitri, Lisa Passarello, Alberto Sardella, Laura Messina, Manuela Pennisi, Daniela Giordano, Simone Palazzo, and Concetto Spampinato. Magnifying facial micro-movements for cognitive evaluation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024 [246]

8. Raffaele Mineo, Isaak Kavasidis, Federica Proietto Salanitri, Lisa Passarello, Giovanni Piccininno, Vincenzo Masciale, Alessandro Anselmo, Cristiano Convertino, Domenico Rotondi, Nicola Laurieri, et al. Automated moca score estimation using eye-gaze data and vision transformers. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 963–968. IEEE, 2025 [242]

9. Salvatore Calcagno, Raffaele Mineo, Daniela Giordano, and Concetto Spampinato. Ensemble and personalized transformer models for subject identification and relapse detection in e-prevention challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023 [39]

10. Raffaele Mineo, Gaia Caligiore, Concetto Spampinato, Sabina Fontana, Simone Palazzo, and Egidio Ragonese. Sign language recognition for patient-doctor communication: A multimedia/multimodal dataset. In *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 202–207. IEEE, 2024 [241]

11. (*) Gaia Caligiore, Raffaele Mineo, Concetto Spampinato, Egidio Ragonese, Simone Palazzo, and Sabina Fontana. Multisource approaches to italian sign language (LIS) recognition: Insights from the multimedalis dataset. In Felice DellOrletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 132–140, Pisa, Italy, December 2024. CEUR Workshop Proceedings [41]

*The candidate contributed equally and should be considered first author.

12. Raffaele Mineo, Gaia Caligiore, Federica Proietto Salanitri, Isaak Kavasidis, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Radar-based imaging for sign language recognition in medical communication. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2025 [240]

13. Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4894–4902, 2025 [249]

14. (*) Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, Jihwan Moon, Václav Javorek, et al. The signeval 2025 challenge at the iccv multimodal sign language recognition workshop: Results and discussion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2025 [222]

15. Raffaele Mineo. Advancing healthcare with multimodal ai: Enhanced diagnosis, cognitive monitoring, and inclusive medical communication. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025 Doctoral Symposium*, volume 16489 of *Lecture Notes in Computer Science*. Springer, 2027 [237]

Patent applications and patent documents

1. (*) Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Procedimento e sistema di elaborazione di immagini angiografiche, 2026. IT Patent 202400014434A1 [78]

2. (*) Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Angiographic image processing method and system, 2025. WO Patent 2025262634A1 [77]

*The candidate contributed equally and should be considered first author.

3. Raffaele Mineo, Paolo Soda, Simone Palazzo, and Concetto Spampinato. Biometric safety system and method for estimating a user hazard condition and triggering an emergency stop, 2027. IT Patent [247]

Additional peer-reviewed publications

1. Alessia Spitalleri, Isaak Kavasidis, Vincenzo Cartelli, Raffaele Mineo, Francesco Rundo, Simone Palazzo, Concetto Spampinato, and Daniela Giordano. Biotrak: A blockchain-based platform for food chain logistics traceability. In *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*, pages 105–110. IEEE, 2023 [343]

2. Federica Proietto Salanitri, Giovanni Bellitto, Raffaele Mineo, Matteo Pennisi, Amelia Sorrenti, Salvatore Calcagno, Daniela Giordano, Simone Palazzo, and Concetto Spampinato. Dynamic graph attention: unraveling spatio-temporal synchrony in eeg data. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2201–2206. IEEE, 2023 [320]

3. Robin Faro, Francesco Cancelliere, Alesssandro Strano, and Raffaele Mineo. Timely predicting power mosfet failure in ev: Ai based metodology and deployment at the edge. In *2025 IEEE Symposia on Computational Intelligence for Energy, Transport and Environmental Sustainability (CIETES Companion)*, pages 1–5. IEEE, 2025 [97]

4. Raffaele Mineo, Amelia Sorrenti, Robin Faro, Gabriele Mineo, Francesco Cancelliere, and Alberto Faro. Pcb-said: A low-cost camera-based dataset for few-shot smd assembly inspection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1351–1357, 2025 [250]

5. Raffaele Mineo, Simone Palazzo, Concetto Spampinato, and Francesco Rundo. Hc-ptq: Poincare-based hyperbolic clustering for data-free quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3994–4001, 2025 [243]

6. Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. A benchmark for radar-based italian sign language recognition using frequency-domain range-time maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026 [248]

7. Ahmed Abul Hasanaath, Raffaele Mineo, Hamzah Luqman, Sarah Alyami, Maad Alowaifeer, Amelia Sorrenti, Gaia Caligiore, Sabina Fontana, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, Motaz Alfarraj, Mufti Mahmud, Simone Palazzo, and Nour Imane Zeghib. Signeval 2026 challenges results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026 [135]

Edited proceedings

1. Simone Palazzo, Giovanni Bellitto, Raffaele Mineo, Amelia Sorrenti, Camila Gonzalez, Angelo Porrello, Lorenzo Bonicelli. Proceedings of the International Workshop on Personalized Incremental Learning in Medicine (PILM '25). ACM SIGMM, 2025 [277]

Others

1. Raffaele Mineo, Paolo Spadaro, Flavio Cannavò, Simone Palazzo, Andrea Cannata, and Concetto Spampinato. Eurovolc-ml/volund-toolkit: Volund v1.0, 2021 [252]

2. Simone Palazzo, Raffaele Mineo, Concetto Spampinato, Flavio Cannavò, Andrea Cannata, Chiara Montagna, Jean-Marie Saurel, O. Geber, and Arnaud Lemarchand. Machine learning tool kit and database. Deliverable report d9.3, EUROVOLC Project, 2021. Horizon 2020 EUROVOLC, Grant Agreement No. 731070 [278]

Awards

A portion of the work summarized in this dissertation was also presented and discussed within the MICCAI 2025 Doctoral Consortium [237], where my doctoral project was accepted for the publication of an official report. Based on the reviewers' evaluation, the submission ranked third in the overall assessment. This experience provided an additional opportunity to position the thesis contributions within the broader MICCAI community and to refine the research framing based on expert feedback.

In addition, the PeRCeiVe Lab submissions to the ICASSP 2023 e-Prevention Signal Processing Grand Challenge achieved 1st place in the relapse detection track on the official hidden test set; this contribution is described in Chapter 4.5.

During the final stage of the doctoral program, I was also awarded an ICCV 2025 Broadening Participation Grant and an ACM Multimedia 2025 Student Travel Grant, supporting participation in two major international conferences in computer vision and multimedia, respectively.

Collectively, these recognitions provided further opportunities to position the thesis contributions within broader research communities and to refine the overall framing based on external feedback.

Statement of Author Contributions and Co-authorship

This dissertation follows a publication-based format. The chapters in Parts I-III are grounded in peer-reviewed papers produced during the doctoral period, and this statement summarizes my contributions to each publication included in the dissertation. Unless otherwise noted, I contributed to (i) the conception of the technical approach, (ii) implementation and experimentation, (iii) analysis and interpretation of results, and (iv) drafting and revising the manuscript.

Peer-reviewed publications

Coronary Artery Stenosis Assessment in X-Ray Angiography Through Spatio-Temporal Attention

I designed and implemented the spatio-temporal modeling pipeline for angiographic sequences, performed preprocessing and training, ran experiments and comparisons, and led the writing of the paper. Clinical collaborators supported problem formulation and interpretation of functional assessment endpoints, while co-authors contributed to reviewing and refining the manuscript.

Non-invasive Physiological Assessment of Intermediate Coronary Stenoses from Plain Angiography: the STARFLOW System

I contributed to the AI methodology and system implementation used for analysis of plain angiography, supported the experimental evaluation and result interpretation, and contributed to drafting and revising the AI-related sections of the manuscript. Clinical co-authors led patient recruitment and clinical protocol definition; all authors reviewed and approved the final manuscript.

A Convolutional-Transformer Model for FFR and iFR Assessment From Coronary Angiography

I co-developed the proposed convolutional-transformer architecture and the experimental protocol, implemented major components of the training and evaluation pipeline, and performed extensive experimentation and analysis. I contributed substantially to writing and revising the manuscript, including method description, experimental sections, and discussion. Co-authors contributed to clinical supervision, dataset collection/annotation, and manuscript revision.

Learning Joint Text and Visual Tokens in CLIP for Medical Image Analysis

I conceived the integration of learnable text and visual tokens for CLIP adaptation in medical imaging, implemented the training pipeline, conducted experiments and ablations, and led the writing. Co-authors contributed to experimental feedback, interpretation, and manuscript revision.

Conditional Visuo-Textual Prompt Learning for Medical Image Analysis

I co-designed the conditional visuo-textual prompt learning framework, implemented the approach, conducted the experimental evaluation and ablations, and drafted major portions of the manuscript. Co-authors contributed to methodological feedback, experimental design refinement, and manuscript revision.

FeDETR: A Federated Approach for Stenosis Detection in Coronary Angiography

I conceived the methodological direction for privacy-preserving stenosis detection in a federated setting, implemented the core training and evaluation pipeline, conducted the experiments and ablation studies, analyzed the results, and led the manuscript writing. Co-authors contributed to data curation and clinical/technical feedback, and participated in manuscript revision.

Magnifying Facial Micro Movements for Cognitive Evaluation

I designed and implemented the computational pipeline for facial micro-movement magnification and analysis, performed experiments and statistical evaluation, and led the writing of the method and results. Co-authors contributed to study design, data collection, clinical interpretation, and manuscript revision.

Automated MoCA Score Estimation Using Eye-Gaze Data and Vision Transformers

I designed the modeling approach for eye-gaze-based automated MoCA estimation, implemented the vision-transformer-based pipeline and temporal encoding, conducted experiments and ablations, and led the writing of the manuscript sections describing the method and results. Co-authors contributed to data collection, protocol definition, and clinical interpretation, and participated in manuscript revision.

Ensemble and personalized transformer models for subject identification and relapse detection in e-prevention challenge

I contributed to the definition and refinement of the methodological pipeline for both challenge tracks, including the overall modeling strategy, the design choices for the ensemble-based subject identification approach, and the formulation of relapse detection as personalized anomaly detection with subject-wise training and score aggregation.

Sign Language Recognition for Patient-Doctor Communication: A Multimedia/Multimodal Dataset

I conceived the dataset design and acquisition strategy for sign language in medical communication, contributed to the sensor setup and data collection protocol, performed baseline experiments, and led the manuscript writing. Co-authors contributed to hardware support, linguistic/clinical guidance, and manuscript revision.

Multisource Approaches to Italian Sign Language (LIS) Recognition: Insights from the MultiMeDaLIS Dataset

I contributed to dataset curation and preprocessing, supported the design and execution of baseline experiments across multiple modalities, and contributed to the analysis and

manuscript writing. Co-authors contributed to linguistic framing, experimental design, and manuscript revision.

Radar-Based Imaging for Sign Language Recognition in Medical Communication

I conceived and implemented the radar-based recognition pipeline, including representation preprocessing and sequence modeling, conducted experiments and comparisons, and led the writing of the technical sections. Co-authors contributed to radar acquisition setup, dataset collection, experimental feedback, and manuscript revision.

Text-Aligned Radar-Based Sign Language Recognition for Healthcare Communication

I co-developed the text-aligned radar-based recognition framework, implemented the contrastive alignment and training pipeline, conducted experiments and ablations, and led drafting of the method and results sections. Co-authors contributed to data collection, experimental feedback, and manuscript revision.

The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and Discussion

I contributed to the definition of the ISLR track based on *MultiMeDaLIS*, including dataset preparation and evaluation protocol, supported the organization and benchmarking activities, and contributed to analysis and writing of the results and discussion sections related to the challenge.

Advancing Healthcare with Multimodal AI: Enhanced Diagnosis, Cognitive Monitoring, and Inclusive Medical Communication

I authored the MICCAI 2025 Doctoral Consortium paper, including conception of the unifying thesis narrative, synthesis of contributions, and writing of the manuscript.

Patent documents

Metodo di diagnosi/prognosi real-time mediante AI in ambito cardiologico

I contributed to the definition of the AI components, system architecture, and experimental evidence supporting the invention, and provided technical input for drafting and refinement of the patent documentation.

Angiographic image processing method and system

I contributed to the technical definition and refinement of the AI pipeline and system claims, supported the preparation of supporting figures/experiments, and provided input during drafting and revision of the international/European extension.

A cardiac biometric P.P.E. as emergency stop device in industrial contexts

I contributed to the conception and development of the invention's core technical solution by defining the AI components and the overall system architecture, and designed and conducted the experimental validation supporting the patent claims, and also provided primary technical input for the drafting, technical review, and refinement of the patent documentation.

Additional notes

All publications were co-authored in accordance with standard academic practice. Each manuscript was reviewed and approved by all listed co-authors prior to submission and publication.

Copyright and Permissions

This dissertation is submitted in a publication-based format. Parts of the content in Chapters 3-5 are based on previously published peer-reviewed articles and are incorporated here in adapted form (e.g., restructured text, unified notation, expanded methodological details, and additional discussion), with appropriate citation to the original sources. In some cases, figures and tables are reproduced or adapted from the corresponding publications.

All third-party material included in this dissertation (including reproduced or adapted figures, tables, and extended excerpts) is acknowledged through proper citation. Where required by the publisher, permissions have been obtained prior to inclusion, and any specific copyright notices requested by the publisher are provided in the relevant figure captions or section footnotes.

Unless explicitly stated otherwise:

- The original text, organization, and presentation in this dissertation are the author's work
- Figures and tables created for this dissertation are the author's work
- Material adapted from the publications listed in the List of Publications is cited using standard academic referencing conventions

Publisher permissions and reuse

The following reuse patterns are applied throughout the dissertation:

- **IEEE publications:** Content is reused in accordance with IEEE policies on author reuse, typically as revised text and re-drawn or properly attributed figures.
- **Springer publications:** Content is reused under Springer's author reuse permissions for theses/dissertations, with proper citation.

- **Oxford University Press (OUP) publication:** The journal article content is summarized and expanded in thesis form, with citation. Any reproduction of figures or tables, if present, is handled according to the relevant OUP permissions requirements
- **CEUR Workshop Proceedings:** Content is reused with citation and, where applicable, attribution for reproduced or adapted figures

Patent documents

This dissertation also references patent documents listed in the List of Publications. Patent texts and figures are used for scholarly discussion of the underlying technical concepts and system design. In Chapter 3.7, which discusses ongoing patenting activities, disclosure is intentionally limited to avoid revealing confidential or non-public details beyond what is already publicly available.

Disclaimer

The views expressed in this dissertation are those of the author and do not necessarily reflect the views of co-authors, affiliated institutions, publishers, or patent offices.

Contents

Abstract i

Lay Summary iii

Preface v

List of Scientific Contributions vii

Awards xiii

Statement of Author Contributions and Co-authorship xv

Copyright and Permissions xxi

Contents xxiii

List of Tables xxxi

List of Figures xxxvii

List of Abbreviations xliii

Acknowledgements xlvii

Dedication xlix

1 Introduction 1

1.1 Research Motivation 1

1.2 Context: AI in Health and Life Sciences 4

1.3 Thesis Objectives and Organization 6

2 Background and Methods **9**
2.1 Deep Learning for Medical Video and Time Series 9
2.2 Transformers in Vision 11
2.3 Multimodal Learning 14
2.4 Foundation Models and Parameter-Efficient Adaptation 17
2.5 Privacy-Preserving and Distributed Learning 20
2.6 Evaluation, Calibration, and Clinical Considerations 22

I AI for Cardiovascular Diagnostics **27**

3 Clinical Background and Challenges in Coronary Artery Assessment **29**

4 Related Work on Coronary Stenosis Evaluation **33**

5 Coronary Artery Stenosis Assessment in X-Ray Angiography for Non-Invasive Evaluation **39**
5.1 Overview 39
5.2 Method 41
5.2.1 Patient cohort, reference standard, and labeling 41
5.2.2 Angiography selection and preprocessing 42
5.2.3 Model architecture 42
5.2.4 Training objective 44
5.2.5 Implementation and optimization protocol 44
5.2.6 Clinical covariates extension and optional late fusion 44
5.2.7 Evaluation protocol and endpoints 45
5.3 Experimental Results 45
5.3.1 Dataset and cohort construction 45
5.3.2 Baseline characteristics 46
5.3.3 Evaluation protocol and metrics 47
5.3.4 Classification performance and comparison with prior methods 47
5.3.5 Regression performance and runtime 50
5.3.6 Impact of clinical covariates 50
5.3.7 Comparison with multi-view clinical baselines and robustness across acquisition systems 52
5.3.8 Key-frame ablation and attention interpretability 53
5.4 Discussion 54
5.4.1 STARFLOW in clinical practice 57

5.4.2 Limitations and future directions 57
5.5 CO_2 Emissions Related to Experiments 58

6 A Convolutional-Transformer Model for FFR and iFR Assessment from Coronary Angiography 59
6.1 Overview 60
6.2 Method 61
6.2.1 The proposed method 61
6.2.2 Feature extraction through 3D convolutions 62
6.2.3 Factorized spatio-temporal transformers 63
6.2.4 Multi-branch prediction 64
6.3 Experimental Results 65
6.3.1 Dataset 65
6.3.2 Training and Evaluation Procedure 67
6.3.3 Effect of input data modality 69
6.3.4 Comparison with the state of the art 70
6.3.5 Ablation Studies 72
6.4 Discussion 76
6.5 Technology Transfer and Intellectual Property 79
6.6 CO_2 Emissions Related to Experiments 80

7 Learning Joint Text and Visual Tokens in CLIP for Medical Image Analysis 81
7.1 Overview 81
7.2 Method 83
7.3 Experimental Results 86
7.3.1 Training and Evaluation Procedure 87
7.3.2 Ablation Studies 96
7.4 Discussion 101
7.5 CO_2 Emissions Related to Experiments 103

8 FeDETR: a Federated Approach for Stenosis Detection in Coronary Angiography 105
8.1 Overview 105
8.2 Method 107
8.2.1 Problem formulation 107
8.3 Experimental Evaluation 109
8.3.1 Dataset 109

8.3.2 Training procedure . . . 109
8.3.3 Results . . . 110
8.4 Discussion . . . 112
8.5 CO_2 Emissions Related to Experiments . . . 113

9 A Cardiac Biometric P.P.E. as Emergency Stop Device 115
9.1 Evaluation, Calibration, and Clinical Considerations . . . 116
9.2 System Architecture . . . 116
9.3 Signal Acquisition, Preprocessing, and Signal-Quality Control . . . 116
9.4 Feature Extraction and Calibration . . . 116
9.5 Dual-Channel Anomaly Detection and Decision Fusion . . . 116
9.6 Safety Chain, Actuation Logic, and Industrial Deployment . . . 116
9.7 Privacy, Logging, and Auditability . . . 116
9.8 Validation Targets, TRL Roadmap, and Standards Alignment . . . 116
9.9 Discussion, Limitations, and IP Disclosure Boundaries . . . 116

II AI-Driven Cognitive and Behavioral Monitoring 117

10 Overview of Automated Cognitive Assessment 119

11 Related Work in Cognitive Monitoring 123

12 Magnifying Facial Micro-movements for Cognitive Evaluation 127
12.1 Overview . . . 127
12.2 Method . . . 128
12.2.1 Architecture Overview . . . 128
12.2.2 Visual Pre-processing . . . 129
12.2.3 Cognitive State Estimation . . . 131
12.3 Experimental Results . . . 132
12.3.1 Dataset . . . 132
12.3.2 Training and Evaluation . . . 133
12.4 Discussion . . . 135
12.5 CO_2 Emissions Related to Experiments . . . 137

13 Automated MoCA Score Estimation Using Eye-Gaze Data and Vision Transformers 139
13.1 Overview . . . 139

13.2 Method . . . 140
13.2.1 The proposed method . . . 140
13.2.2 Gaze-Conditioned Patch Selection . . . 141
13.2.3 Time-Aware Positional Embedding (TAPE) . . . 142
13.2.4 Transformer-Based Gaze Encoding . . . 143
13.2.5 MoCA Score Prediction . . . 144
13.3 Evaluation . . . 145
13.3.1 Dataset acquisition . . . 145
13.3.2 Experimental setting . . . 147
13.3.3 Performance . . . 147
13.4 Discussion . . . 148
13.5 CO_2 Emissions Related to Experiments . . . 151

14 Ensemble and personalized transformer models for subject identification and relapse detection 153
14.1 Introduction . . . 153
14.2 Methods . . . 154
14.2.1 Track 1 - Ensemble for subject identification . . . 154
14.2.2 Track 2 - Personalized relapse detection via reconstruction-based anomaly scoring . . . 156
14.3 Results . . . 158
14.4 Discussion . . . 160
14.5 CO_2 Emissions Related to Experiments . . . 163

III Sign Language Recognition for Medical Communication 165

15 Importance and Challenges in Automatic SLR 167

16 Related Work of SLR 171

17 Sign Language Recognition for Patient-Doctor Communication: A Multimedia/Multimodal Dataset 175
17.1 Overview . . . 175
17.2 Data acquisition and methodology . . . 176
17.2.1 Selection of Signs . . . 176
17.2.2 Selection of Tools for Data Collection . . . 177
17.2.3 The Recording Protocol . . . 181

17.2.4 Discussion on Usability and Reliability 182
17.3 Discussion 183

18 Multisource Approaches to Italian Sign Language (LIS) Recognition: Insights from the MultiMeDaLIS dataset 185
18.1 Overview 185
18.2 Testing the Dataset 186
18.3 Models and Architectures 187
18.4 Training and Evaluation Procedure 188
18.5 Results 188
18.6 Discussion 190

19 Radar-Based Imaging for Sign Language Recognition in Medical Communication 193
19.1 Overview 193
19.2 Method 195
19.2.1 Dataset Description 195
19.2.2 Data Preprocessing 195
19.2.3 Model Architecture 196
19.2.4 Training procedure 198
19.3 Experimental Results 199
19.4 Ablation Study 200
19.4.1 Ablation on Autoencoder Reconstruction 200
19.4.2 Observations on Skip Connections and Dropout 200
19.4.3 Ablation on Different Backbone Configurations 201
19.4.4 Ablation on RDM Spectrum Threshold 201
19.4.5 Ablation on Backbone Type and Pretraining 202
19.5 Discussion 203
19.6 CO_2 Emissions Related to Experiments 204

20 Text-Aligned Radar-Based Sign Language Recognition for Healthcare Communication 205
20.1 Overview 205
20.2 Method 207
20.2.1 Problem formulation 207
20.2.2 Text encoding 208
20.2.3 RDM encoding and classification 209

20.3 Experimental evaluation . . . 209
20.3.1 Dataset . . . 209
20.3.2 Data preprocessing . . . 210
20.3.3 Metrics . . . 211
20.3.4 Training procedure . . . 211
20.3.5 Results . . . 211
20.3.6 Ablation study . . . 213
20.4 Discussion . . . 215
20.5 CO_2 Emissions Related to Experiments . . . 217

21 Impact and Community Contribution: The SignEval 2025 Challenge at the ICCV Multimodal SLR Workshop 219
21.1 Overview . . . 219
21.2 Track 1: Continuous Sign Language Recognition Track . . . 220
21.2.1 Dataset . . . 221
21.2.2 Evaluation Metrics . . . 222
21.2.3 Results and Participating Teams . . . 222
21.3 Track 2: Isolated Sign Language Recognition . . . 226
21.3.1 Task Description . . . 226
21.3.2 Dataset . . . 226
21.3.3 Evaluation Metrics . . . 227
21.3.4 Results and Participating Teams . . . 227
21.4 Discussion . . . 230

22 Conclusion and Future Work 233
22.1 Cardiovascular Diagnostics: Multi-View Angiography and Clinical Trials . . 233
22.2 Cognitive and Behavioral Monitoring: Multimodal Fusion and Real-Time Deployment . . . 235
22.3 Sign-Language Recognition: Real-Time Clinical Integration and User Studies 237

List of Tables

5.1 Baseline features: STEMI, ST-segment elevation myocardial infarction; NSTEMI, non ST-segment elevation myocardial infarction. 48
5.2 Comparison of our method with state-of-the-art general deep learning and clinical AI methods 49
5.3 Performance of STARFLOW for overall patients, FFR patients, and iFR patients. 50
5.4 Performance in terms of MSE and MAE when regressing FFR and iFR values 50
5.5 Comparison of STARFLOW with state-of-the-art methods. 52
5.6 Sensitivity analysis: performance of STARFLOW across various X-ray systems. 52
5.7 Effect of key-frame integration. 53

6.1 Classification performance comparison of our model using different input data modalities. 67
6.2 Per-hospital classification performance obtained by the proposed approach. Performance is consistent across the five centers despite the high data imbalance. 69
6.3 Classification performance comparison with the state of the art. For each model, we report the input modality as well as balanced accuracy (*Acc.*), AUC, sensitivity (*Sens.*) and specificity (*Spec.*), computed through 5-fold nested cross-validation. S stands for single view, i.e., only one view per patient is used, while M stands for multiview, i.e., at least two views per patient are employed. 69
6.4 Calibration performance in terms of Brier score; lower values indicate better calibration. 72
6.5 Ablation on architecture modules. Results refer to classification performance. 73
6.6 Ablation on FFR-iFR regression on a subset of the overall dataset. Results refer to classification performance. 73
6.7 Ablation on number of frames. Results refer to classification performance. . . 74
6.8 FFR and iFR regression performance of the proposed model. 74
6.9 Ablation on attention strategies. 76

7.1 Performance comparison with different combinations of textual and visual tokens on the FFR 2-class TVTL-CLIP. *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer. 90

7.2 Performance comparison with different combinations of textual and visual tokens on the FFR 2-class CTVTL-CLIP (ShuffleNet V2 support). *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer. 91

7.3 Performance comparison with different combinations of textual and visual tokens on the FFR 3-class TVTL-CLIP. *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer. 92

7.4 Performance comparison with different combinations of textual and visual tokens on the FFR 3-class CTVTL-CLIP (ShuffleNet V2 support). *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer. 93

7.5 Comparison with SOTA architectures on the FFR 2-class benchmark. "# params" reports trainable weights in millions. For each backbone we list results when trained from scratch and when finetuned from an ImageNet initialization (grey rows). All metrics are percentages; Acc. reports balanced accuracy. TVTL-CLIP and CTVTL-CLIP are shown using their best configuration on this benchmark. 94

7.6 Comparison with SOTA architectures on the FFR 3-class benchmark. "# params" reports trainable weights in millions. For each backbone we list results when trained from scratch and when finetuned from an ImageNet initialization (grey rows). All metrics are percentages; Acc. reports balanced accuracy. TVTL-CLIP and CTVTL-CLIP are shown using their best configuration on this benchmark. 95

7.7 Text-only ablation study on the FFR 2-class benchmark. Performance is reported when no visual prompts are used. "*Text Tok.*" denotes the number of learnable textual prompt tokens, while "*CSC*" marks whether class-specific context is applied or not. TVTL and CTVTL are reported side-by-side; since CSC is not varied in CTVTL, the same values are repeated across CSC rows. All metrics are expressed in %. 97
7.8 Text-only ablation study on the FFR 3-class benchmark. Performance is reported when no visual prompts are used. "*Text Tok.*" denotes the number of learnable textual prompt tokens, while "*CSC*" marks whether class-specific context is applied or not. TVTL and CTVTL are reported side-by-side; since CSC is not varied in CTVTL, the same values are repeated across CSC rows. All metrics are expressed in %. 98
7.9 Results using only learnable visual tokens on the FFR 2-class benchmark (no textual prompts). (*Vis. Tok.* number of visual tokens, *Deep Vis. Tok.* w/wo layer-wise visual tokens). TVTL and CTVTL are reported side-by-side. . . . 98
7.10 Results using only learnable visual tokens on the FFR 3-class benchmark (no textual prompts). (*Vis. Tok.* number of visual tokens, *Deep Vis. Tok.* w/wo layer-wise visual tokens). TVTL and CTVTL are reported side-by-side. . . . 99
7.11 Ablation on different CLIP visual backbones on the FFR 2-class benchmark, using the fixed prompt setup of TVTL (Text Tok.=64, CSC=True, Vis. Tok.=5, Deep Vis. Tok.=Yes). For CTVTL, the same token setup is used (CSC is fixed by design in the available runs). 100
7.12 Ablation on different CLIP visual backbones on the FFR 3-class benchmark, using the fixed prompt setup of TVTL (Text Tok.=64, CSC=True, Vis. Tok.=5, Deep Vis. Tok.=Yes). For CTVTL, the same token setup is used (CSC is fixed by design in the available runs). 100

8.1 Performance comparison between the proposed approach, FedAvg and FedBN, for different combinations of number of rounds and epochs, on the target metrics. Standard deviations refer to inter-node variability. 111

12.1 MMSE prediction accuracy using various combinations of task data (T), emotions (E) and visual pre-processing techniques, including face stabilization and cropping (FSC), video motion magnification (VMM), and MHI. 134
12.2 Individual question prediction accuracy using various combinations of task data (T), emotions (E) and visual pre-processing techniques. 134
12.3 MMSE prediction accuracy at different MHI granularities. 135

13.1 Game-MoCA attribute associations. VS: Visuo-spatial, N: Naming, AT: Attention, L: Language, AB: Abstraction, DR: Delayed recall, O: Orientation. . 146
13.2 Performance comparison for different MoCA-related tasks. 147
13.3 Performance comparison when different positional embedding methods are used. 148
13.4 Performance comparison using different time unit durations. 148

14.1 Aggregated feature set for Track 2 (computed on 5-minute slices and stacked over the day) [427]. 155
14.2 Track 1 ensemble configurations used in the submission. "PE" denotes positional embedding. A duplicated configuration is assigned weight 2 in the ensemble. 156
14.3 Track 2 - Best per-subject configurations selected by grid search. "Aggr." indicates aggregated representation, "Raw" indicates raw representation. . . 158
14.4 Performance comparison among participant teams on the provided test set for subject identification. We report the raw test accuracy for each team. Our score ranks 2nd. 158
14.5 Performance comparison among participant teams on the provided test set for relapse detection. We report the ROC-AUC, the PR-AUC and the harmonic mean of the previous scores, for each team. Our score ranks 1st. 159
14.6 Performance of tested models on validation set for each subject. We reported the harmonic mean between ROC-AUC and PR-AUC. For CNN and VolUnD we tested also *raw* and *aggregated* data. 159
14.7 Track 1 - Input-feature ablation on the CNN baseline (validation). "Sleep" indicates inclusion of sleeping status; "Steps" indicates step-derived signals; "Calories" indicates calories per time unit 160
14.8 Track 1 - Window size study (validation). Reported values are the final train and validation losses and accuracies for each window length 160
14.9 Track 2 - VolUnD validation summary over the hyperparameter search (maximum validation score across runs). 161

17.1 Labels of the selected signs grouped by category 178
17.2 Recording tools and corresponding acquired data 180

18.1 Performance of SL-GCN multi-stream on the test set. 189
18.2 Comparison of different datasets on SL-GCN model. 189
18.3 Performance of various methods on the *MultiMeDaLIS* Dataset. 190

19.1 Summary of existing datasets, in terms of data modality, number of signs, number of subjects and number of repetitions per sign. Notation: for RADAR-only data modality, we report the sensor frequency; for "signs", we report by default the number of sign-language signs, as well as letters ("L") or non-sign-language gestures ("G"). * Plus RGB-D and face tracking data. . . 196

19.2 Comparison with state-of-the-art methods in sign language recognition. The "Modality" column indicates the primary sensor modality employed by the method. 199

19.3 Autoencoder reconstruction error under different conditions. Skip connections consistently reduce MSE, with larger gains as training progresses and as more data are used. 200

19.4 Impact of skip-dropout in the autoencoder backbone, with skip connections always enabled. A high dropout level of 0.8 yields the best overall balance and the highest classification performance. 201

19.5 Effect of different RDM spectrum thresholds. The -90 dB setting provides the best trade-off between suppressing background noise and preserving discriminative motion energy. 202

19.6 Comparison of backbone architectures and training strategies (mean $\pm$ std). Freezing consistently improves generalization, and our residual autoencoder achieves the best overall performance 203

20.1 Comparison with state-of-the-art methods in sign language recognition. The "Data type" column indicates the primary sensor modality employed by the method. The best performance is achieved by TRACE, which combines 3×RDM and 3×MTI inputs, highlighting the importance of multimodal radar encoding and multi-perspective information. TRACE leverages the MixConv architecture pretrained on radar data and employs 64 soft prompts, yielding the best overall performance among all methods. 212

20.2 Impact of radar input types and number of antennas on model performance. We evaluate the effect of using RDM, MTI, and their combination, in both single-antenna and multi-antenna configurations. The best performance is achieved by combining 3×RDM and 3×MTI inputs, highlighting the importance of multimodal radar encoding and multi-perspective information. 213

20.3 Effect of backbone architecture and pretraining on performance. We compare ResNet 3D [134], R(2+1)D [364], and MixConv [354] with and without pretraining on radar data. Results show that both architectural choice and pretraining substantially impact performance, with the pretrained MixConv achieving the highest scores across all metrics. 214
20.4 Effect of the number of soft prompts on classification performance. We evaluate the effect of varying the number of learnable prompts prepended to each class label. Results show that using 64 prompts leads to the best performance across all metrics, highlighting the benefit of richer prompt-based conditioning for textual supervision. 215

21.1 Results for Task 1 (Signer-Independent) in SignEval 2025 Challenge Track 1, showing both development-set and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams. 223
21.2 Results for Task 2 (Unseen Sentence) in MSLR Challenge Track 1, showing both development-set and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams. 224
21.3 Baseline and Leaderboard for the SignEval Challenge Track 2. 228

List of Figures

5.1 Two sample angiography views for the same patient. Red bounding boxes show the major stenosis. 40

5.2 Architecture overview. A 3D CNN, common to all views, extracts spatio-temporal features which are subsequently refined through a multi-head spatio-temporal self-attention. The resulting features are then channeled to a three-output layer which carries out binary classification (over the threshold), and quantification of iFR and FFR. 43

5.3 Patients excluded from the dataset used. 46

5.4 Distribution of percentage of stenosis across the final cohort. 47

5.5 ROC (left) and precision-recall (right) curves comparing our approach with state-of-the-art clinical AI methods. 49

5.6 AUC-ROC (left) and precision-recall curve on a fold (right) for all patients. . 50

5.7 AUC-ROC (above) and precision-recall curve on a fold (below) for FFR patients (left) and for iFR patients (right). 51

5.8 AUC-ROC (above) and precision-recall curve on a fold (below) for comparing FFR/iFR datasets (left) and comparing SOTA methods (right). Central illustration: the STARFLOW model. 53

5.9 Impact of attention strategies. Interpretability maps computed through M3D-CAM [51] with attention enabled vs disabled. Yellow denotes the most activated regions, purple the least activated ones. With attention, activations concentrate on vessels and stenotic segments; without attention, activations extend to non-coronary anatomy visible in X-rays. 54

6.1 Architecture of the proposed approach for stenosis significance assessment. Input angiography videos are first processed by a pre-trained 3D convolutional model for local feature extraction. Then, self-attention layers based on transformers are employed to capture intra-relations in space and time, and the resulting intermediate representation is fed to three output branches, providing predictions as either a significance class or a regression estimate of FFR/iFR values. 62

6.2 Examples of the two views per subject. Bounding boxes indicate major stenoses as identified by two expert radiologists. 66

6.3 Cumulative confusion matrix over the five tested folds. 67

6.4 ROC (left) and Precision-Recall (right) curves comparison between our approach and state-of-the-art methods. 71

6.5 Interpretability maps, computed by M3D-CAM [119, 51], of our method (last row) and DMQCA [410], HEAL [409] and DMTRL [396] (first, second, and third row, respectively). For each image, we also report, as a red bounding box, the location of the major stenosis, as identified by cardiologists. In each map, the yellow parts are the most activated ones, while the purple parts are the least activated ones. 72

6.6 MAE at different FFR and iFR range intervals. Blue bars measure the regression MAE computed at each interval of the FFR/iFR ranges. Standard deviation on each interval is reported as an error bar. Vertical red lines correspond to thresholds for FFR and iFR, below which a stenosis is considered to be hemodynamically significant. 75

6.7 (First row) Bland-Altman plot between the predicted and the reference values for FFR and iFR. There is agreement between the two sets of values with 0 mean, suggesting no systematic bias in the evaluation. (Second row) Scatter plots for FFR and iFR demonstrate that the lowest errors and the smaller standard deviations are consistently observed around clinically-relevant thresholds (0.8 for FFR, 0.89 for iFR, depicted in blue dashed lines). 75

6.8 Attention mechanisms. *Global attention*: each spatio-temporal location (in blue) attends to all other locations (in green). *Temporal attention*: each location (in blue) attends the same spatial location along the time dimension (in green). *Factorized spatio-temporal attention*: attention is first computed on each frame independently, and then along the temporal dimension over frame-level features. 77

6.9 Interpretability maps, computed through M3D-CAM [119, 51], when using different attention strategies. For each image, we also report, as a red bounding box, the major stenosis as identified by clinicians. In each part, yellow regions correspond to the most activated areas, whereas purple regions correspond to the least activated ones. 77

7.1 TVTL-CLIP architecture. The text encoder receives learnable tokens and class labels to produce a text embedding for each class. The image encoder (ViT-B/32) receives learnable visual tokens and patch embeddings to produce a visual embedding. Classification is performed by maximizing image-text similarity for the correct class 84

7.2 CTVTL-CLIP architecture. A support network produces an instance-specific context token that modulates the learnable prompts. The conditioned text prompts and class labels are encoded by the frozen text encoder, while the conditioned visual prompts and patch embeddings are encoded by the frozen vision encoder. The model predicts by image-text similarity, as in CLIP . . . 85

8.1 Each node trains a DETR model using its local dataset, composed of key frames extracted from coronary angiography videos. At the beginning of each round, the central server sends the aggregated backbone to each node, which uses it to extract features from the key frames. After each round, each node sends its locally trained backbone back to the central server, which aggregates them all. 107

8.2 Given a key frame, the CNN backbone learns a 2D representation, which is then flattened by the model. Before passing these learned features to the transformer encoder, the model also extends them with positional encoding. The transformer decoder outputs some learned positional embeddings and feeds them into a feed-forward network that predicts classes and their corresponding bounding boxes. 108

8.3 Figure 8.3a shows representative examples of *low severity stenoses* (illustrated in the top row) and *mild or high severity stenoses* (depicted in the bottom row). Figure 8.3b presents an averaged heatmap, which is derived from the collective bounding box positions across all samples. 110

12.1 Overview of the proposed approach. Facial micro-movement features are extracted through a visual pre-processing pipeline and combined with emotional states and performance metrics on cognitive tasks to predict a subject's MMSE score 128
12.2 Examples of motion history images at different temporal granularities. The three columns represent the effects using windows of 1, 10 and 60 seconds. The top row shows the videos after face stabilization and cropping, and the bottom row shows the same videos after motion magnification 131

13.1 DA-ViT (Dynamic-Aligned Vision Transformer) architecture. Our major contribution lies in replacing the standard positional encoder with TAPE, which incorporates information about the temporal duration of fixation point. This helps standardize the representation of time, compensating for the highly variable eye-gaze sampling rates, which can differ significantly across consecutive time intervals. 141
13.2 Example screenshots from the cognitive games presented to the subjects. From the top left: *Complete the sequence*, *Word search*, *Coffee preparation*, *Order by size*, *At the grocery shop*, and *Bill payment*. 145

17.1 Data acquisition: (a) RGB, (b) depth map; (c) infrared stereoscopic with laser projection (left vision), (d) infrared stereoscopic with laser projection (right vision). 179
17.2 Range doppler map from the three radar antennas. 181
17.3 Face (a) and body (b) tracking output. 181
17.4 Background color variation to differentiate the preparation (yellow) and recording (green) phases of the sign. 182

19.1 Overview of the end-to-end architecture of the proposed method. 197
19.2 Example reconstructions after 10 epochs. Skip connections preserve fine motion-energy structures that are attenuated when relying on the bottleneck alone 200
19.3 Qualitative comparison of RDMs obtained with different spectrum thresholds 202

20.1 Overview of the *TRACE* method. Radar-based RDM sequences are encoded into visual embeddings $\mathbf{z}_i^v$ using a 3D CNN. Each class label is represented textually and prepended with learnable prompts before being processed by a frozen text encoder, yielding text embeddings $\mathbf{z}_k^t$. A contrastive loss $\mathcal{L}_{\text{con}}$ encourages alignment between matching visual–text pairs in the shared embedding space. 207

20.2 Overview of the inference phase. The CNN encodes the RDM sequence into a visual embedding $\mathbf{z}_i^v$, which is compared with all text embeddings $\mathbf{z}_k^t$ corresponding to class labels. The predicted class is the one with the highest cosine similarity between $\mathbf{z}_i^v$ and $\mathbf{z}_k^t$. 208
20.3 Overview of the radar signal preprocessing pipeline leading to the RDM. The diagram shows the main stages of the processing chain, including range FFT, Doppler FFT, and thresholding. While simplified, the figure highlights the key transformation from raw time-domain radar data to a 2D RDM representation. Intermediate operations such as mean removal, windowing, zero-padding, and FFT shifting are omitted for clarity. 211
20.4 RDM with and without MTI filtering. Comparison between the same RDM without MTI filtering (left) and with MTI filtering applied (right). The MTI filter suppresses static background signals by zeroing out bins below a predefined velocity threshold, resulting in a cleaner representation of dynamic hand motion. 212

List of Abbreviations

General AI/ML

2D+t	Two-Dimensional + Time
3D-CNN	Three-Dimensional Convolutional Neural Network
AI	Artificial Intelligence
ANN	Artificial Neural Network
CL	Contrastive Learning
CNN	Convolutional Neural Network
DL	Deep Learning
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
KD	Knowledge Distillation
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
NCE	Noise-Contrastive Estimation
OOD	Out-of-Distribution
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SOTA	State of the Art
SSL	Self-Supervised Learning
XAI	Explainable Artificial Intelligence

Multimodal/Representation Learning

ASR	Automatic Speech Recognition
CLIP	Contrastive Language-Image Pretraining
CoCoOp	Conditional Context Optimization
CoOp	Context Optimization
IMU	Inertial Measurement Unit

LoRA	Low-Rank Adaptation
mAP	Mean Average Precision
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
RGB	Red-Green-Blue
RGB-D	Red-Green-Blue and Depth
TTS	Text-to-Speech
ViT	Vision Transformer
VLM	Vision-Language Model
VPT	Visual Prompt Tuning

Evaluation/Metrics

ACC	Accuracy
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
F1	F1-score
FPR	False Positive Rate
MAE	Mean Absolute Error
MSE	Mean Squared Error
PR	Precision-Recall
ROC	Receiver Operating Characteristic
TPR	True Positive Rate

Clinical/Cardiovascular

CAD	Coronary Artery Disease
CAG	Coronary Angiography
DS	Diameter Stenosis
FFR	Fractional Flow Reserve
iFR	Instantaneous Wave-Free Ratio
LAD	Left Anterior Descending artery
LCx	Left Circumflex artery
PCI	Percutaneous Coronary Intervention
QCA	Quantitative Coronary Angiography
RCA	Right Coronary Artery

Cognitive/Behavioral Monitoring

AFib	Atrial Fibrillation
ECG	Electrocardiogram
HR	Heart Rate
HRV	Heart Rate Variability
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
PPG	Photoplethysmography

Sign Language/Accessibility

ASL	American Sign Language
CSLR	Continuous Sign Language Recognition
HoH	Hard of Hearing
ISLR	Isolated Sign Language Recognition
LIS	Lingua dei Segni Italiana
SLR	Sign Language Recognition

Radar/Sensing

FMCW	Frequency-Modulated Continuous Wave
mmWave	Millimeter-Wave
MTI	Moving Target Indication
RA	Range-Angle
RD	Range-Doppler
RF	Radio Frequency
SNR	Signal-to-Noise Ratio

Systems/Privacy/Safety Engineering

DP	Differential Privacy
E-stop	Emergency Stop
FL	Federated Learning
GDPR	General Data Protection Regulation
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
MPC	Multi-Party Computation
PL	Performance Level
PLC	Programmable Logic Controller
PPE	Personal Protective Equipment

SIL	Safety Integrity Level
TRL	Technology Readiness Level
V&V	Verification and Validation

Hardware/Embedded/Reliability

AFE	Analog Front-End
CMRR	Common-Mode Rejection Ratio
ESD	Electrostatic Discharge
GPU	Graphics Processing Unit
LED	Light-Emitting Diode
MCU	Microcontroller Unit
NPU	Neural Processing Unit
SQI	Signal Quality Index
TPU	Tensor Processing Unit

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their guidance, support, and trust throughout my Ph.D. journey. I am grateful to Prof. Concetto Spampinato, Prof. Simone Palazzo, Prof. Francesco Rundo, and Prof. Danilo Avola for their mentorship, for encouraging both scientific depth and practical impact, and for providing an environment where ambitious ideas could be explored and refined.

I am sincerely thankful to Prof. Fabrizio D'Ascenzo and Dr. Ovidio De Filippo (DSM, University of Turin) for their invaluable collaboration on the cardiology projects, which led to the realization of an international patent. My thanks extend to Prof. Daniela Giordano (DIEEI, University of Catania) and Prof. Manuela Pennisi (BIOMETEC, University of Catania) for their contribution to the research on cognitive decline and behavioral monitoring. I am grateful to Prof. Sabina Fontana and Dr. Gaia Caligiore (DISUM, University of Catania), and Prof. Egidio Ragonese (DIEEI, University of Catania) for their collaboration on the research in sign language recognition and radar-based sensing. Finally, I am also grateful to Prof. Paolo Soda (University Campus Bio-Medico of Rome) for his trust and support in the technology-transfer activities, including the ongoing patenting process. Their mentorship, expertise, and encouragement have been fundamental in shaping this dissertation.

I would also like to thank my collaborators and co-authors, whose expertise and commitment made this work possible. I am particularly grateful to the clinical partners involved in the cardiovascular studies for their essential contributions to data collection, study design, and interpretation of results, and to the colleagues and collaborators who supported the cognitive monitoring and sign language recognition lines of research with significant feedback, data acquisition efforts, and interdisciplinary perspectives.

I express my sincere gratitude to the University Campus Bio-Medico of Rome and the University of Catania for providing the academic environment and resources. I also thank the laboratories, administrative staff, and technical personnel for their support.

Finally, I would like to thank my family and colleagues for their constant encouragement and patience. Their support has been essential throughout the challenges and milestones of this Ph.D., and I am deeply grateful for their presence and understanding.

To my father and to **my mother**, for their love, strength, and constant presence.

To my grandparents, **to my aunts and uncles**, and **to my cousins**, for the roots you gave me and for your steady support.

To my supervisor *Concetto*, now Full Professor, who has been my mentor and who, despite my missteps, has always shown generosity, given me many second chances, and continued to welcome me into his lab.

To my supervisor *Simone*, who more than anyone else has been my "sponsor", for his concrete, daily trust and support.

To my closest colleagues and daily co-conspirators in academia *Fede*, *Giovannino*, and *Matteuccio*, for making the university journey lighter, funnier, and far less lonely, and for being there with help, feedback, and perfectly timed irony whenever it mattered most.

And finally, **to humanity**, in the hope that this small drop of work may contribute, even if only slightly, to making knowledge more useful and technology more human.

"If I have seen further, it is by standing on the shoulders of giants." — Isaac Newton

Chapter 1

Introduction

1.1 Research Motivation

Artificial intelligence is increasingly used for medical decision support, patient monitoring, and assistive technologies, yet translation into routine practice remains uneven [360, 299, 180, 271]. A recurring reason is that real-world impact is often constrained less by the expressiveness of the learning architecture than by the feasibility of acquiring the right input data at scale [180, 271]. In healthcare, acquisition is frequently expensive, invasive, scarce, privacy-sensitive, or operationally incompatible with everyday workflows [180]. As a result, even technically strong models may remain confined to settings where acquisition conditions are ideal, data are abundant, and deployment constraints are relaxed [180, 408].

This dissertation is motivated by the observation that, in many clinically relevant problems, the clinical question is clear, but the most informative measurements are not consistently accessible [180, 271]. Coronary physiology is a representative example: while invasive indices such as fractional flow reserve (FFR) and instantaneous wave-free ratio (iFR) provide guideline-supported evidence for functional lesion relevance [267, 183, 288, 325], their routine use can be limited by procedural burden, costs, and operational inertia [267, 183]. Similar acquisition frictions appear in longitudinal cognitive assessment, where repeated in-person testing is hard to scale, despite the clinical role of standardized instruments such as the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA) [105, 265]. They also arise in clinical communication with Deaf patients, where well-known barriers in access and interaction can persist in time-critical settings [151]. In these scenarios, improving performance under laboratory conditions is necessary but not sufficient; what matters is whether the overall pipeline can be executed reliably with feasible sensing, appropriate privacy safeguards, and a deployment footprint compatible with real clinical constraints [180, 408].

The central motivation of this thesis is therefore to treat the input modality and the acquisition pipeline as first-class design variables, rather than as fixed preconditions [180, 271]. The guiding principle is to keep strong modeling paradigms while leveraging multimodal signals from complementary modalities. Instead of exclusively optimizing architectures under a fixed sensing setup, this thesis asks how the same clinical objectives can be pursued using inputs that reduce barriers to adoption, while preserving clinical relevance [180, 271]. Concretely, the dissertation studies how spatio-temporal attention models and Transformers [373], vision transformers [85], contrastive vision-language alignment [295], and parameter-efficient adaptation [148] can be paired with acquisition-aware choices such as angiography video analysis without manual keyframe/vessel intervention, commodity vision signals and wearable time series for behavioral monitoring, and privacy-preserving sensing (including radar-based modalities and federated training settings) when data sharing is constrained [231, 175, 314, 126]. The goal is not to replace clinical gold standards, but to expand the set of feasible pathways for screening, triage, monitoring, and communication support.

Three cross-cutting requirements shape this motivation throughout the dissertation:

- **Accessibility**: enabling clinically meaningful inference from inputs that are less invasive, less costly, and more widely available across sites and populations
- **Privacy**: reducing reliance on identifiable modalities and enabling learning and inference under privacy constraints, including distributed settings where data sharing is limited [231, 175, 314]
- **Deployability**: designing end-to-end pipelines that are robust to real-world variability, compatible with operational workflows, and feasible under computational and latency constraints [299, 408]

These requirements motivate multimodal learning as the unifying methodological lens of the thesis, and provide a coherent thread across its three application parts [20, 269]. Healthcare signals are intrinsically heterogeneous: imaging sequences, physiological time series, behavioral cues, text, and sensor measurements often provide complementary evidence about the same underlying state [20, 269]. Multimodal learning enables principled fusion and redundancy, improves robustness under partial observability, and supports semantic grounding when raw signals are ambiguous [20, 269, 295]. At the same time, multimodal systems raise the challenges that this dissertation explicitly addresses: alignment across modalities and time, calibration under domain shift and multi-site variability [408], robustness to missing or unreliable channels [20, 269], and evaluation under clinically meaningful operating points rather than only aggregate statistical metrics [180, 266, 183].

The three research tracks are therefore connected by a common systems-level problem: in healthcare, the clinically most informative signal is often not the signal that can be acquired most easily, repeatedly, safely, or privately. Cardiovascular diagnostics, cognitive and behavioral monitoring, and sign-language recognition instantiate this problem under different constraints. In cardiovascular diagnostics, the objective is to approximate functional information that would otherwise require invasive pressure-wire measurements, starting from routinely acquired angiographic sequences and, when possible, complementary views, procedural context, or distributed multi-center evidence. In cognitive monitoring, the objective is to obtain repeated assessment from low-burden behavioral and wearable signals when specialist-administered testing is difficult to scale. In medical sign-language recognition, the objective is to preserve accessible communication while balancing recognition accuracy with privacy and clinical acceptability through RGB, depth, radar, inertial streams, and text-based semantic grounding. Thus, the three tracks are not only linked by the use of multimodal models, but by a shared design principle: complementary sensing is used to reduce acquisition barriers while keeping the clinical objective fixed.

Within this framing, the dissertation develops three application domains that exemplify distinct acquisition barriers and corresponding opportunities for multimodal design. In cardiovascular diagnostics, the focus is on physiology-informed assessment from coronary angiography, grounded in the clinical role of FFR/iFR for intermediate lesions and revascularization decision-making [266, 183, 359, 71, 118]. In cognitive and behavioral monitoring, the objective is scalable screening from non-invasive behavioral signals captured with commodity devices and consumer wearables, enabling longitudinal assessment outside specialized clinics [105, 265]. In sign language recognition for medical communication, the emphasis is inclusive interaction in healthcare, where privacy-preserving sensing and robust multimodal grounding are essential for realistic adoption [184, 305, 151, 126], including scenarios where standard RGB capture is undesirable. Across these domains, the methodological emphasis is consistent: the choice and combination of modalities are engineered to reduce acquisition burden while keeping the clinical target fixed.

Overall, the motivation of this thesis is to contribute methods and systems that do not treat data acquisition as an afterthought [180, 271]. Instead, it argues that meaningful progress in medical AI requires an explicit coupling between learning models and the realities of sensing, privacy, and deployment [180, 408]. By following the principle of "keep strong modeling paradigms while leveraging multimodal signals from complementary modalities", and by using multimodality as the connecting thread across diagnosis, monitoring, and accessibility [20, 269], the dissertation aims to broaden the practical reach of AI in healthcare [360, 180].

1.2 Context: AI in Health and Life Sciences

Over the past decade, advances in machine learning have reshaped the technical landscape of health and life sciences [93, 210]. Deep learning is now routinely used to extract clinically meaningful information from medical images, physiological signals, electronic health records, and multimodal patient assessments [210, 132, 301, 332, 20]. This progress has been enabled by three converging factors: (i) broader availability of digital health data and improved acquisition devices [228, 346], (ii) the maturation of scalable neural architectures, from convolutional networks to transformers [193, 373, 85], and (iii) the emergence of foundation models pre-trained on large corpora that can be adapted to downstream tasks with limited supervision [34, 295, 37].

However, healthcare remains a distinctive operating environment that differs from conventional machine perception settings [180, 271, 138]. The most informative measurements are often invasive, costly, or operationally burdensome, creating an intrinsic tension between clinical validity and scalability [180, 271]. Modalities that maximize information content may be impractical for frequent monitoring or broad deployment, whereas widely available modalities may be noisier, less standardized, and more variable [180, 138]. Clinical data are also heterogeneous across patient populations, acquisition devices, protocols, and sites, inducing domain shift and challenging reproducibility [408, 351, 104]. At the same time, privacy and governance constraints limit data sharing and complicate the creation of centralized datasets, particularly for identifiable modalities such as video and high-resolution imaging [231, 175, 314, 292]. Finally, clinical decision-making is threshold-driven and risk-driven: success is not defined only by average accuracy, but by sensitivity, specificity, calibration, robustness, and interpretability at clinically meaningful operating points [64, 122].

Within this landscape, multimodal learning has gained importance because many clinical questions are not fully resolved by a single signal [20, 269, 20]. Patient state is expressed across complementary sources of evidence in both space and time: anatomical imaging captures structure, dynamic imaging and physiological measurements capture function, behavioral signals reflect cognitive and affective processes, and language mediates documentation, decision support, and accessibility [20, 269, 20]. Multimodal learning provides a principled framework to integrate these signals through fusion, alignment, and cross-modal reasoning, improving robustness when one channel is degraded or unavailable and supporting generalization under real-world variability [20, 269, 20]. Equally important, multimodality is not only a modeling choice but an acquisition strategy: by combining weaker, more scalable modalities, one can often approach the clinical utility of stronger but less feasible measurements [20].

In parallel, transformers and attention mechanisms have influenced medical AI in two

complementary ways [373, 85, 19, 327]. First, attention-based architectures provide a flexible means to model long-range temporal structure in video and time series [373, 19]. which is central to dynamic clinical processes such as contrast propagation in angiography, temporal trajectories in behavioral monitoring, and sequential dynamics in sign language. Second, token-based representations support parameter-efficient adaptation and reuse of pre-trained knowledge [148, 147, 194], which is valuable when labeled medical data are scarce [210, 138], when distribution shifts across sites are common [408, 351, 104], and when updates must be performed with limited supervision and constrained compute budgets [147, 194]. This paradigm also aligns naturally with multimodal settings, where attention and tokenization provide a shared interface for aligning heterogeneous streams [34, 295, 37].

A further contextual trend is the increasing emphasis on privacy-preserving learning and deployment [175, 314, 175]. Federated learning, secure computation, and privacy-aware sensing are becoming practical necessities in many healthcare scenarios [231, 175, 314, 175, 174]. Rather than treating privacy as a post-hoc constraint, modern systems increasingly incorporate it as a design variable that shapes modality selection, training strategy, and inference pipelines [175, 314, 175]. Deployment considerations (compute efficiency, latency, robustness to acquisition variability, and integration into clinical workflows) have similarly become central determinants of whether AI systems can move beyond proof-of-concept studies [180, 408, 138]. These requirements are particularly salient when models are intended to operate across institutions, on constrained hardware, or in settings where data capture and sharing are tightly regulated [231, 175, 314].

This dissertation is situated at the intersection of these trends, but it uses them to impose a clearer organizing logic across application domains. Instead of treating cardiovascular diagnostics, cognitive monitoring, and medical sign-language recognition as unrelated verticals, the thesis frames them as instances of the same systems problem: achieving clinically meaningful inference under acquisition, privacy, and deployment constraints by leveraging multimodality and transformer-based modeling [180, 20, 269, 175, 314]. In cardiovascular diagnostics, the work focuses on angiography-based assessment of functional stenosis relevance and consolidates the end-to-end system contribution within the cardiovascular part to preserve a coherent translation pathway [266, 183]. In cognitive and behavioral monitoring, the work studies scalable assessment from commodity vision signals that can support longitudinal screening outside specialized clinics [105, 265]. In sign language recognition for medical communication, the work addresses inclusive interaction under privacy constraints, including the use of alternative sensing that reduces identifiability while preserving semantic content [151, 126]. Across these domains, the consistent message is that robust clinical AI depends not only on the model class, but on how modeling choices co-evolve with feasible multimodal

acquisition pipelines, privacy requirements, and deployable system design [180, 408, 20, 269].

1.3 Thesis Objectives and Organization

The overarching objective of this dissertation is to advance multimodal learning methods and end-to-end systems that broaden the practical reach of AI in healthcare. The thesis is grounded in a recurring observation: in many clinically relevant problems, progress is limited less by the lack of expressive models than by the difficulty of acquiring informative inputs at scale, under privacy constraints and within realistic workflows. Accordingly, the dissertation pursues four tightly connected goals: (i) reduce input acquisition barriers while preserving the clinical objective, by enabling inference from routine, low-cost, or privacy-preserving signals; (ii) develop robust spatio-temporal and multimodal models that operate under heterogeneous acquisition conditions, domain shift, and imperfect supervision; (iii) treat privacy and deployability as first-class requirements, encompassing both training paradigms (e.g., federated learning) and sensing choices (e.g., non-identifiable modalities) as well as computational footprint and integration constraints; and (iv) provide evidence through reproducible evaluation, including datasets, baselines, and benchmarking protocols aligned with clinically meaningful operating points (e.g., sensitivity/specificity trade-offs and threshold-driven decisions).

In line with this acquisition-aware perspective, the thesis delivers a set of contributions across three domains (cardiovascular diagnostics, cognitive/behavioral monitoring, and medical sign language recognition) and distills reusable methodological components for multimodal learning under constrained sensing and supervision. Across the dissertation, multimodality is used both to integrate complementary evidence and to redesign acquisition pipelines when canonical inputs are invasive, costly, privacy-sensitive, or operationally misaligned with routine practice.

The organization of the dissertation follows this acquisition-aware interpretation of multimodality. The three application parts should therefore be read as complementary case studies of the same problem rather than as independent verticals. Part I addresses a setting in which the desired clinical quantity is functional, but the deployable input is routinely acquired angiography. Part II addresses a setting in which longitudinal assessment is clinically useful, but repeated specialist testing is difficult to scale; the thesis therefore explores low-burden behavioral and wearable signals. Part III addresses a setting in which communication accessibility requires visual-gestural understanding, but identifiable video may be inappropriate in sensitive clinical environments; the thesis therefore studies multimodal and radar-based privacy-preserving alternatives. Across these cases, the common objective is to design learning

pipelines in which modality choice, fusion, and substitution are part of the technical solution.

Chapter 1 introduces the research motivation, the contextual landscape of AI in health and life sciences, and the acquisition-aware principle that connects the thesis.

Chapter 2 provides methodological background. It reviews deep learning for medical images, video, and time series; transformer-based architectures and attention mechanisms for temporal reasoning; multimodal fusion and alignment; foundation models and parameter-efficient adaptation (including prompt learning); privacy-preserving and distributed learning with emphasis on federated optimization under non-IID data; and clinical evaluation aspects such as calibration, threshold-based decision-making, and robustness across sites.

Part I develops cardiovascular methodologies and a translational system view from routine coronary angiography: spatio-temporal learning for anatomical and functional assessment, foundation-model adaptation under domain shift with prompt-based vision-language modeling, federated learning for stenosis assessment (FeDETR), and deployment-oriented considerations (*CardioAngel*), including integrated pipeline and technology-transfer aspects.

Part II develops non-invasive cognitive and behavioral monitoring pipelines from commodity sensing, combining facial micro-dynamics analysis, transformer-based eye-gaze modeling, and long-term wearable time-series analysis to support scalable assessment under uncontrolled acquisition, irregular sampling, and missing data.

Part III develops multimodal sign language recognition for medical communication under privacy and accessibility constraints, integrating dataset/benchmarking contributions, privacy-preserving mmWave radar-based recognition, text-aligned semantic grounding, and community benchmarking studies, with a closing discussion on deployment feasibility.

Chapter 22 concludes the dissertation and outlines future research directions across the three parts, with emphasis on multi-center validation and prospective studies in cardiology, robust multimodal fusion and real-time deployment for behavioral monitoring, and real-world clinical integration and user studies for sign-language-based medical communication.

Chapter 2

Background and Methods

2.1 Deep Learning for Medical Video and Time Series

Many clinically relevant phenomena are inherently temporal [98]. Contrast propagation in angiography, physiological rhythms in biosignals, gaze trajectories during cognitive tasks, and motion patterns in gestural communication are examples where the diagnostic or behavioral signal is distributed across time rather than concentrated in a single snapshot. Deep learning for medical video and time series therefore aims to (i) extract informative spatial representations at each time instant, (ii) model temporal dependencies and dynamics, and (iii) remain robust under real-world acquisition constraints such as variable frame rates, noise, missing segments, and heterogeneous protocols across sites [98, 408].

Temporal healthcare data support multiple learning objectives. In **classification**, the goal is to predict discrete categories (e.g., stenosis present/absent, impairment risk class, sign label). In **regression**, the objective is to estimate continuous clinical indices (e.g., functional scores such as FFR/iFR, cognitive screening scores). In **detection** and **localization**, models may output structured predictions such as bounding boxes, key frames, or temporal segments [177, 363]. In practice, supervision is frequently **weak** or **sparse**: labels may be available only at exam level, frame-level annotations may be impractical, and outcomes can be noisy or threshold-driven (e.g., clinically meaningful cut-offs) [152]. This motivates architectures that can aggregate evidence over time, support calibrated decision-making at specific operating points [123], and when needed provide interpretable evidence about which portions of the sequence drove the prediction [324].

Early deep learning solutions for sequential biomedical data relied heavily on convolutional designs because of their efficiency and inductive bias. For physiological signals (e.g., ECG, PPG), **1D CNNs** provide local temporal feature extraction and can be stacked to expand

receptive fields, offering stable optimization and high parallelism compared to recurrent alternatives [181, 302]. For video, a common baseline is **2D CNNs with temporal aggregation**, where a 2D backbone encodes frames independently and features are aggregated by pooling, temporal averaging, or lightweight temporal modules; this approach is computationally efficient and often effective when supervision is sequence-level or when discriminative information is largely spatial [177]. **3D CNNs** extend kernels across space and time to jointly capture appearance and short-term motion, and they have been widely adopted in medical video settings (including ultrasound and angiography) when relevant signals lie within local temporal windows [363, 48]. Their main limitations are the computational footprint and the difficulty of capturing long-range dependencies without deep stacks, aggressive downsampling, or clip-based processing [363, 48]. In generic video understanding, **two-stream and optical-flow-based** designs explicitly separate appearance and motion [338]; in medical settings, explicit motion signals can be useful when subtle temporal changes are diagnostically informative, but their reliability depends on acquisition conditions and image quality, and they can be sensitive to artifacts and compression [408].

Sequence-modeling approaches were introduced to handle longer temporal context and variable dynamics. **RNNs, LSTMs, and GRUs** [24, 145, 58] iteratively update a hidden state and can be applied directly to time-series samples or combined with frame-wise CNN features for video. In practice, however, recurrent models can be sensitive to long sequences, irregular sampling, and optimization stability, and they have increasingly been complemented or replaced by attention-based designs [373]. **Temporal convolutional networks (TCNs)** [192, 19] provide an alternative based on stacks of dilated 1D convolutions, yielding long effective receptive fields with stable gradients [19]. When designed with causal convolutions, TCNs naturally support streaming and bounded latency, making them attractive in deployment settings where real-time constraints matter [192, 19].

A recurring theme in modern medical video analysis is the combination of local feature extraction with global dependency modeling [363, 373, 327]. Hybrid designs are motivated by the observation that medical videos exhibit both local signals (e.g., vessel edges, textures, contrast boundaries) and global spatio-temporal structure (e.g., contrast filling patterns across the vessel tree). **Temporal attention** mechanisms allow models to weight frames or segments by relevance, which is useful in weakly supervised scenarios where only sequence-level labels are available [373, 152]. **Spatio-temporal attention** extends this idea by modeling relationships among distant spatial regions across time, supporting phenomena where clinically meaningful information is not localized and must be integrated across multiple frames [384]. **CNN-Transformer hybrids** operationalize this combination by using CNN/3D CNN modules to generate compact feature maps and then applying transformer blocks to model global

interactions among tokenized representations, balancing inductive bias and computational efficiency with long-range reasoning [373, 26, 327].

Real-world medical temporal data often violate assumptions commonly made in standard video benchmarks, and robustness must be designed explicitly [408]. **Variable frame rate and irregular sampling** arise across devices and sites, requiring resampling strategies, time-aware embeddings, or architectures tolerant to temporal irregularity [98]. **Limited sequence length and clipping** are common because many pipelines operate on clips extracted from longer procedures; clip selection can be automatic or operator-dependent, so models should tolerate variability in boundaries and content [177, 363]. **Noise, compression, and motion artifacts** can strongly affect temporal signals, motivating quality-aware preprocessing and augmentation strategies [408]. Finally, **domain shift across sites and devices** can alter both appearance and temporal dynamics, which motivates multi-center evaluation, calibration analysis, and, when centralization is infeasible, privacy-preserving collaborative learning [123, 408, 231, 175, 174].

The methods in this section provide the foundations for the contributions developed in later parts of the dissertation. In Part I, spatio-temporal modeling is central to interpreting angiographic sequences for stenosis severity and functional relevance under heterogeneous acquisition. In Part II, temporal reasoning supports the analysis of behavioral dynamics from facial video and eye-gaze traces acquired with commodity devices. In Part III, sequence models capture the dynamics of sign execution, including privacy-preserving radar representations. Across these domains, the thesis leverages complementary strengths of convolutional feature extractors and attention-based sequence modeling, while maintaining a consistent focus on acquisition-aware robustness and deployability.

2.2 Transformers in Vision

Transformers have reshaped modern computer vision by replacing fixed, locality-driven aggregation (typical of convolutions) with content-adaptive attention [373, 85, 215]. This shift is particularly relevant in healthcare, where clinically meaningful information may be distributed across spatially distant regions, unfold over time, or manifest through subtle patterns that are difficult to capture with purely local operators [327]. In addition, the growing heterogeneity of acquisition devices and protocols in clinical practice makes robustness to domain shift and variability a central requirement [408]. This section summarizes the core concepts of vision transformers, the main design directions introduced to improve efficiency and spatio-temporal modeling, and the practical considerations that motivate their use in this dissertation [327].

At the core of the transformer architecture is **self-attention** [373], which computes interactions among a set of tokens. In vision, tokens can represent image patches, region proposals, feature vectors produced by a convolutional stem, or modality-specific descriptors (e.g., radar-derived maps treated as image-like grids) [85, 215]. By modeling token-to-token dependencies, self-attention enables the network to focus on semantically relevant relationships, for instance between vessel segments, between facial regions whose micro-dynamics co-vary over time, or between spatial structures that appear with different contrast or noise levels [327]. A standard transformer block combines multi-head self-attention (capturing diverse interaction patterns in parallel), a position-wise feed-forward network (FFN/MLP) that refines each token representation, and residual connections with normalization (commonly LayerNorm) to support stable optimization [373].

Because attention is permutation-invariant, transformers require explicit **positional information** to preserve spatial layout and temporal order [373, 85]. In vision, this is typically handled through absolute positional embeddings (learned or sinusoidal), relative positional embeddings that encode pairwise spatial offsets, or attention biases designed for window-based attention [328, 215]. In spatio-temporal settings, positional encoding often extends to time, either by adding temporal embeddings to frame tokens or by defining factorized space-time positional schemes [26, 14]. These choices have practical implications: they influence the model's ability to generalize across resolutions, support variable sequence lengths, and maintain robustness when sampling is irregular or when frames are missing [26, 14].

The **Vision Transformer (ViT)** [85] introduced a simple yet effective formulation: an image is split into fixed-size patches, each patch is linearly projected to an embedding token, and the token sequence is processed by a stack of transformer blocks. A special class token can be prepended to aggregate global information for classification [85]. ViT benefits from global receptive field at every layer, flexible context modeling, and strong scalability with data and model size [85]. In medical imaging, however, labeled data are often limited, and training large transformers from scratch can be unstable or prone to overfitting [362, 327]. As a result, pre-training and adaptation become central: medical ViT pipelines typically rely on large-scale initialization (supervised, self-supervised, or vision-language pre-training) and then employ careful regularization, parameter-efficient tuning, or task-specific token strategies to adapt to clinical domains [362, 47, 139, 295, 327].

A direct application of global self-attention has quadratic complexity in the number of tokens, which can be prohibitive for high-resolution images and long sequences [373, 382]. Several efficiency-oriented transformer variants address this issue. One prominent family uses **window-based attention**, exemplified by Swin Transformer [215], which restricts attention to local windows and shifts windows across layers to enable cross-window interactions while

building hierarchical representations [215]. Other approaches rely on **hierarchical tokenization** by progressively merging or pooling tokens to reduce sequence length at deeper layers, supporting multi-scale reasoning [215, 391]. A further set of methods approximates attention through **sparse** patterns or **low-rank** factorization, trading exactness for computational savings, with effectiveness depending strongly on hardware and implementation [382, 59]. In healthcare, these efficiency choices are not only about training cost: they also determine whether inference can run within the latency and memory budgets compatible with deployment, and whether models can process realistic input sizes without aggressive downsampling that might discard clinically relevant detail [327].

For video and temporal vision tasks, transformers can be extended in several ways [26, 14]. A straightforward strategy is to extract spatial tokens per frame and apply attention across both space and time, enabling the model to integrate evidence over the sequence [26]. To reduce complexity, many designs use **factorized space-time attention**, where spatial attention is applied within frames and temporal attention is applied across frame-level representations or across corresponding patch tokens [26, 14]. Alternatively, **spatio-temporal tokenization** can encode short temporal windows or local motion into tokens before attention is applied [14]. These approaches are relevant for angiography and behavioral monitoring, where contrast dynamics, gaze trajectories, and micro-movements carry key information, and they also transfer naturally to radar-based sign language recognition when radar representations are treated as temporal sequences of image-like maps [327, 26].

A core modeling trade-off in medical settings is **inductive bias versus flexibility**. Compared to CNNs, transformers impose weaker locality and translation-invariance priors [85, 362]. This can be advantageous when relevant dependencies are long-range or non-local, but it can be harmful when supervision is scarce and local structure is essential [362, 327]. Hybrid designs often provide a pragmatic balance: CNN-Transformer architectures exploit convolutional inductive bias for local feature extraction and use attention for global integration [391, 68], while domain-informed token choices (e.g., ROI tokens, vessel-centric tokens, or modality-specific tokens) inject structure without hard-coding a rigid pipeline [327]. This balance aligns with the broader acquisition-aware perspective of the dissertation: selecting inductive bias is part of making strong models operate reliably under constrained and variable inputs [327].

Finally, clinical decision support requires attention to **calibration, interpretability, and reliability**. Transformers provide tools such as attention maps, token saliency, and gradient-based attributions that can help inspect model behavior [53], but attention is not inherently an explanation and must be validated against domain knowledge and failure modes [158]. Similarly, transformer models can be overconfident under domain shift; therefore

evaluation should emphasize operating-point metrics (sensitivity, specificity), calibration analysis, and when possible multi-center or external validation rather than relying solely on aggregate accuracy [123, 275, 408]. In this dissertation, attention-derived signals are used primarily as diagnostic instruments to understand model behavior, while reliability is addressed through task-appropriate metrics and acquisition-aware evaluation protocols [327].

Transformers appear throughout the dissertation in complementary roles. In Part I, attention-based components and token-centric representations support spatio-temporal modeling of angiographic sequences and contribute to functional assessment from plain angiography. In Part II, transformer-based sequence modeling supports behavioral inference from eye-gaze and facial dynamics under commodity acquisition. In Part III, attention-based encoders and sequence transformers support radar-based sign language recognition and text alignment for semantic grounding. More broadly, the token perspective links naturally to parameter-efficient adaptation and prompt learning, which are leveraged when pre-trained representations must be adapted to specialized clinical domains with limited supervision and strict deployment constraints [423, 163, 148, 147].

2.3 Multimodal Learning

Clinical phenomena are rarely captured by a single source of evidence [20, 303, 269, 253]. Anatomical structure, physiological function, behavior, and communication jointly shape health outcomes, and they are observed through heterogeneous modalities such as images, videos, biosignals, text, and sensor measurements [20, 303]. Multimodal learning provides principled tools to combine these sources, enabling complementary information extraction, more reliable inference under noise and variability, and graceful degradation when one modality is missing or degraded [20, 303, 269]. In this dissertation, multimodality is not treated as an optional enhancement, but as a practical design lens: when the most informative measurement is invasive, expensive, privacy-sensitive, or operationally inconvenient, a multimodal perspective can preserve the clinical objective by exploiting alternative signals or combinations that are more feasible to acquire and deploy [20, 253].

Multimodality matters in healthcare for several intertwined reasons [20, 253]. First, different modalities offer **complementary** views of the same underlying condition: images and videos capture anatomy and appearance, physiological signals capture function and rhythms, behavioral traces capture cognitive state, and language mediates communication and contextual information [20, 303]. Second, clinical data are often noisy and heterogeneous; combining independent evidence streams can reduce uncertainty and improve reliability, including calibration at clinically meaningful operating points [20]. Third, **partial observability** is

common: modalities may be unavailable due to cost, time, equipment, patient tolerance, or privacy and governance constraints [20, 253]. Fourth, cross-modal constraints can improve generalization by discouraging overfitting to modality-specific artifacts and encouraging representations that are stable under domain shift across devices and sites [20, 303]. These properties are central to the thesis goal of acquisition-aware learning, where the objective is not only to improve predictive performance under ideal conditions, but also to maintain clinically meaningful inference under realistic sensing pipelines [253].

A fundamental multimodal problem is how to combine information from different sources, commonly referred to as **fusion** [20, 15, 341]. A first family of methods performs **early fusion**, combining modalities at the input level or at shallow feature levels [20, 15]. Early fusion can capture low-level correlations but typically requires precise alignment in space and time and can be brittle under missing modalities [20, 52]. A second family uses **late fusion**, where each modality is processed independently and predictions or high-level features are combined near the output, for instance via logit averaging, stacking, or gated ensembles [20, 341, 157]. Late fusion is naturally robust to missing modalities but may underuse fine-grained cross-modal interactions [20]. Between these extremes, **intermediate fusion** combines features at one or more mid-level representations, often providing a good trade-off between interaction modeling and robustness, especially when modalities differ in structure and noise properties [20, 303]. Attention mechanisms further enable **attention-based fusion** and **cross-attention**, where one modality provides queries and the other provides keys/values, allowing the model to learn which cross-modal correspondences matter for the task [373, 367, 219]. This is particularly relevant when one modality carries semantic guidance (e.g., text describing a concept) and the other carries sensory evidence (e.g., images or radar maps) [367, 295].

In practice, clinical modalities are often not naturally synchronized, and alignment cannot be assumed [20, 303]. Gaze trajectories may be sampled irregularly, radar and video streams may have different frame rates and latencies, and clinical metadata may be recorded only at sparse events [336]. Multimodal learning therefore requires alignment strategies that respect acquisition constraints rather than imposing unrealistic synchronization assumptions [20, 336]. Typical solutions include resampling and interpolation to reconcile rates, time-aware embeddings that encode sampling times or temporal gaps, and windowed alignment that couples modalities at a coarser granularity instead of enforcing per-frame correspondence [336, 52]. These considerations are particularly important in behavioral monitoring and alternative sensing, where acquisition is designed for scalability and privacy rather than for controlled laboratory synchronization [253].

Beyond fusion, many multimodal systems aim to map heterogeneous inputs into a **shared latent space** that preserves task-relevant semantics [20, 269]. One approach relies on

joint encoders that process concatenated or fused representations, which can capture strong correlations but often presuppose co-observed modalities and may generalize poorly when modalities are missing [20, 303]. A more flexible approach uses **separate encoders with shared objectives**, where each modality has its own encoder and training enforces consistency between their embeddings [269, 20]. This design supports modality-specific preprocessing, enables inference from a single modality when others are absent, and provides a natural pathway for modality substitution when acquisition constraints change [20].

A dominant paradigm for learning shared spaces is **contrastive representation learning** [127, 273, 56, 140] .Contrastive objectives pull together embeddings of matched pairs (positive samples) and push apart mismatched pairs (negative samples) [127, 273]. In the vision-language setting, CLIP [295] is a canonical example, aligning images and text in a shared embedding space that supports similarity-based inference and prompting. In healthcare, contrastive alignment is attractive because it can be label-efficient and leverage weak supervision (e.g., reports, structured metadata, glosses or textual descriptors), can improve robustness by discouraging reliance on modality-specific artifacts, and can facilitate modality substitution when one channel is unavailable at inference [295, 20, 253]. However, contrastive learning also introduces risks that are particularly salient in clinical contexts: negative sampling can create false negatives when different samples share latent clinical similarity, and alignment alone does not guarantee causal validity or clinical correctness [60]. In this dissertation, contrastive alignment is employed as a pragmatic mechanism for semantic grounding and generalization, complemented by task-specific evaluation and clinically meaningful operating-point analysis [253].

A recurring operational constraint is **missing modalities** [20, 268]. Because it is rarely feasible to collect all modalities for all patients across sites, multimodal systems should be explicitly designed to operate under missing inputs [20, 268]. Practical strategies include modality dropout during training to encourage robustness [268], mixture-of-experts and gating to dynamically weight available modalities [157], and teacher-student or distillation schemes where a rich modality supervises a more deployable modality [144, 124]. Generative imputation can also be used to reconstruct missing modalities, but in clinical settings it must be treated cautiously due to reliability and auditability concerns [253]. The emphasis in this thesis is on deployable pathways: when a modality is not feasible (for privacy, cost, or workflow reasons), the goal is to retain performance using alternative inputs and principled cross-modal supervision rather than reconstructing sensitive data [253].

Multimodal learning connects directly to all three application domains addressed in the dissertation [253]. In cardiovascular diagnostics, multimodal perspectives motivate the integration of temporal imaging evidence with auxiliary context and support generalization

via vision-language alignment and parameter-efficient adaptation [295]. In cognitive and behavioral monitoring, the focus is on extracting clinically relevant information from vision-derived behavioral signals, where robustness to noise, variability, and missingness is essential for scalable assessment [253, 336]. In sign language recognition for medical communication, multimodality is explicit in both data and method: multiple sensor streams support fusion and benchmarking, and privacy-preserving recognition leverages radar as an alternative modality, with text alignment enabling semantic grounding [295, 20]. Across these domains, multimodal learning provides the methodological foundation for the dissertation's central theme: reducing the barrier of input acquisition while preserving clinically meaningful objectives through robust fusion, alignment, and deployable modality choices [20, 253].

2.4 Foundation Models and Parameter-Efficient Adaptation

Foundation models have changed how visual representations are learned and reused [34, 279]. Instead of training task-specific networks from scratch on limited domain data, current pipelines increasingly rely on large-scale pretraining to acquire general-purpose features, followed by lightweight adaptation to downstream problems [279, 82, 141]. This shift is particularly relevant in healthcare, where labeled datasets are often small, heterogeneous across sites, expensive to curate, and sometimes infeasible to centralize due to governance constraints [253, 20]. In such settings, the practical question is not only whether a model can achieve high accuracy in a controlled benchmark, but whether it can be adapted, validated, and deployed reliably with limited supervision, constrained compute, and minimal maintenance overhead [253]. Foundation models and parameter-efficient adaptation provide a coherent answer to these requirements: they separate the cost of learning broad representations from the cost of specializing them to a specific clinical task, site, or protocol [34, 194].

In vision, large-scale pretraining supervised/self-supervised yields backbones that transfer effectively across tasks [82, 141, 85, 47, 139]. Even when pretraining data are non-medical, early and intermediate features often remain useful, while higher layers can be adapted to reflect domain-specific concepts [402, 279]. In vision-language learning, foundation models such as CLIP align images and text in a shared embedding space through contrastive objectives, enabling similarity-based inference and flexible conditioning via language [295, 162]. This is attractive in healthcare because textual supervision may be available at scale through reports, structured findings, or procedure notes, potentially reducing dependence on dense annotations [295, 253]. However, medical language differs substantially from web text:

terminology is specialized, abbreviations are common, and the mapping between words and clinical concepts can be context-dependent [147, 406]. As a result, direct zero-shot transfer frequently underperforms, and adaptation becomes essential both for visual domain shift and for the semantics of the textual label space [388, 147].

While full fine-tuning of large backbones can yield strong performance, its limitations are amplified in clinical practice [253]. First, with scarce supervision, fine-tuning many parameters increases overfitting risk and can produce unstable decision boundaries under domain shift [253, 402]. Second, retraining large models for each task or site can be computationally expensive and difficult to reproduce, especially when hyperparameters and data preprocessing are not perfectly standardized [253]. Third, multi-site deployment often requires modest but persistent calibration or domain adaptation; maintaining a separate fully fine-tuned model per site increases storage, validation, and regulatory burden [253]. Fourth, privacy and governance constraints may require that adaptation be performed locally, with limited parameter updates exchanged or audited [20, 253]. These factors motivate parameter-efficient fine-tuning (PEFT) methods, which adapt a model by updating only a small subset of parameters or by introducing lightweight trainable modules while keeping the foundation backbone largely frozen [194, 163].

Prompt-based adaptation is a particularly natural form of PEFT for vision-language models [423, 147]. In CLIP-like architectures, prompts provide the conditioning context that maps a label space into the text encoder [295, 147]. Handcrafted templates can be effective in natural-image domains but often fail to encode clinical nuance, where a generic word may have a specific meaning (e.g., "erosion") or where labels correspond to subtle visual patterns [388, 147]. Soft prompt tuning addresses this issue by replacing or augmenting discrete prompt text with a small set of learnable tokens in the embedding space [423, 147]. These tokens are optimized for the downstream task while leaving the pretrained weights untouched, which reduces overfitting and supports fast adaptation [423]. Prompting can also be extended to the visual stream through visual prompt tuning, where learnable tokens or perturbations are injected into the image token sequence of a vision transformer [402]. This becomes important when the target domain differs substantially from pretraining: modulating only the textual branch may not be sufficient if the visual encoder does not attend to clinically relevant patterns [388, 402].

A further step is conditional prompting, where prompts are not static but depend on the input instance or on contextual information such as acquisition protocol, device type, or site [422]. In heterogeneous medical datasets, a single shared prompt can be suboptimal because the same clinical label may correspond to diverse appearances, and systematic biases may arise from devices or protocols [253, 20]. Conditional prompting aims to capture this variability while remaining parameter-efficient, by producing sample-

conditioned prompt vectors or by combining learned prompts with context features derived from auxiliary networks [422]. This notion of conditioning is central when robustness to acquisition variability is a primary design goal rather than an afterthought [253].

In addition to prompting, adapter modules provide a complementary PEFT strategy [194]. Adapters insert small trainable bottleneck layers inside transformer blocks or between layers, keeping the backbone largely frozen [194]. They can learn task-specific transformations with a limited parameter budget and support modular deployment by swapping lightweight components rather than duplicating entire backbones [194]. Adapters are also useful in multi-task and continual settings, where new tasks are added over time and catastrophic forgetting must be controlled through restricted updates [194].

Low-Rank Adaptation (LoRA) is another widely used PEFT technique, injecting low-rank update matrices into selected weight tensors, typically attention projections and sometimes MLP layers [148]. LoRA often approaches full fine-tuning performance while training far fewer parameters, making it attractive when compute and storage are constrained [148]. In medical contexts, LoRA can facilitate domain adaptation across sites without duplicating full models and can support distributed training regimes where exchanging compact updates is preferable to exchanging full weights [148, 20]. In this thesis, LoRA provides an important reference point within the broader PEFT landscape, even when the core contributions emphasize prompting mechanisms that explicitly treat tokens as the adaptation interface [423, 402, 147].

Across all PEFT methods, an essential design question is what to freeze and where to adapt [402, 194]. In healthcare, stable adaptation is frequently preferable to aggressive weight updates, particularly when datasets are small and domain shift is present [253]. Common patterns include freezing the backbone and tuning only prompts, adapters, or LoRA parameters; adapting higher layers more than lower layers when the domain shift is moderate; and using regularization, careful validation protocols, and early stopping to avoid spurious improvements [402, 194]. For vision-language models, the choice also involves the text encoder: adapting the label representation through learned tokens can be as important as adapting the visual stream, because clinical label spaces and terminology can be misaligned with generic pretraining corpora [388, 147, 423].

In this dissertation, foundation models and PEFT are leveraged primarily to support generalization under constrained supervision and heterogeneous acquisition [253]. In Part I, CLIP-style representations are adapted through learnable tokens in both modalities and through conditional mechanisms that address variability at the level of the input instance or context, providing a pathway to robust performance without full fine-tuning [295, 423, 402, 422]. More broadly, the PEFT perspective serves the thesis objective of deployability: adapting strong pretrained models with minimal updates reduces computational cost, facilitates

reproducibility, and enables practical personalization or site-specific calibration while keeping system maintenance manageable in real clinical workflows [253, 34, 194].

2.5 Privacy-Preserving and Distributed Learning

Privacy is a first-order constraint in healthcare [309, 2, 175]. Many clinically relevant data streams are inherently identifiable (e.g., video, high-resolution imaging, and longitudinal behavioral signals), and governance requirements often constrain how data can be stored, transferred, and centralized [309, 2, 314]. These constraints shape not only data management practices, but also the feasible design space of AI systems [175]. In practice, privacy limitations are rarely isolated technical details: they interact with acquisition protocols, institutional policies, and deployment workflows, and they frequently determine whether a model can progress from a retrospective study to routine use [175, 314]. This section summarizes the distributed learning paradigms and privacy-preserving strategies most relevant to this dissertation, with emphasis on federated learning under non-IID data, evaluation methodology in multi-center settings, and the relationship between privacy constraints and sensing choices [424, 175].

Healthcare data are sensitive for ethical, legal, and operational reasons, and privacy concerns can limit both cross-institutional sharing and the use of certain modalities [309, 2]. For example, facial video used for behavioral assessment and RGB recordings used for sign language recognition may be unacceptable in real deployments even when consent is obtained, because identifiability persists and secondary-use risks remain [309, 2, 175]. At the same time, clinical datasets are naturally fragmented: they are distributed across hospitals, collected under different acquisition protocols and devices, and curated under site-specific workflows [175, 314, 73]. Two implications follow. First, learning must often be distributed across institutions without transferring raw data [198, 314, 175]. Second, in some scenarios, the input modality itself must be redesigned to reduce identifiability while preserving task-relevant information [297]. These directions align with the dissertation's acquisition-aware perspective: privacy is not only a constraint on learning but also a driver for modality selection [175].

Federated learning (FL) provides a practical framework for collaborative training when centralization is infeasible [198, 424, 175]. In its standard setup, multiple clients (e.g., hospitals) perform local optimization on private datasets and communicate model updates to a coordinating server, which aggregates them into a global model [198]. The most common baseline is FedAvg, where the server averages client updates, typically weighted by local sample size [198]. FL can substantially reduce the need for centralized data pools and can

enable learning from the diversity of multi-center cohorts; however, it should not be conflated with a complete privacy guarantee [424, 175]. Depending on the threat model, model updates can leak information, and additional protection mechanisms may be required [111, 334, 176].

A core difficulty of medical FL is statistical heterogeneity: client data are typically non-IID [424, 175]. Differences arise from demographics and disease prevalence, imaging devices and acquisition protocols, annotation practices, and local inclusion criteria [175, 73, 126]. Non-IID data can induce client drift, slow or destabilize convergence, and yield global models that disproportionately fit dominant sites while underperforming on smaller or systematically different populations [424, 322]. Addressing heterogeneity has therefore become central to practical FL [424]. Representative strategies include regularized objectives such as FedProx, normalization-aware variants such as FedBN, control-variates approaches (e.g., SCAFFOLD) that reduce client drift, client sampling and weighting schemes that mitigate dominance effects, and personalized or clustered FL methods that relax the assumption that a single global model is optimal for all sites [202, 88, 322, 13, 353, 1, 424]. In this dissertation, heterogeneity is treated as an experimental factor with direct clinical relevance rather than as a nuisance, particularly in multi-center cardiovascular applications where protocol variability is unavoidable [73, 126].

In addition to FL, several mechanisms can strengthen privacy or reduce leakage risks [424, 175]. Differential privacy (DP) limits information that can be inferred about any individual by introducing controlled noise, typically to gradients or updates [174, 35]. DP provides formal guarantees but can reduce utility, especially when datasets are small, models are large, or the task requires high sensitivity to subtle patterns [174, 35]. Secure aggregation protocols reduce exposure of per-client updates by ensuring the server observes only an aggregated contribution rather than individual client messages, mitigating risks from honest-but-curious servers or accidental disclosure of site-specific updates [116]. More general secure computation approaches, including multi-party computation and homomorphic encryption, can enable training or inference on encrypted data but may impose substantial overhead for high-dimensional updates and large transformer-based models [112, 123, 254]. In real deployments, algorithmic privacy mechanisms coexist with governance measures such as auditability, access control, data minimization, and explicit agreements on model dissemination; these institutional layers often determine what is practically feasible even when a method is technically sound [309, 2, 175].

A distinctive theme of this thesis is that privacy can be addressed at two complementary levels [175]. Learning-level privacy retains the same modality but modifies how learning is performed, via FL, DP, secure aggregation, or secure computation [198, 174, 116, 112]. Input-level privacy modifies the sensing modality itself to reduce identifiability by design [297].

Input-level privacy is particularly compelling in settings where storing or sharing identifiable recordings is unacceptable even if learning is distributed [309, 2]. For instance, mmWave radar can capture motion patterns relevant to sign language without recording facial identity or background details, providing a deployable pathway in sensitive environments [297]. This perspective connects directly to the thesis-wide principle that acquisition is a first-class design variable: changing the input can relax privacy constraints that would otherwise block deployment [175].

Evaluation under privacy constraints requires methodology that mirrors real deployment conditions rather than relying on pooled and homogenized test sets [73, 126]. In multi-center settings, results should be reported per site and under held-out domains to quantify generalization and identify failure modes [73, 126, 175]. Because clinical decisions are threshold-driven and risk-driven, evaluation should emphasize sensitivity, specificity, and calibration at clinically meaningful operating points, rather than relying solely on overall accuracy [300]. Heterogeneity can also create systematic performance differences across populations or sites, motivating fairness-aware reporting and subgroup analyses when feasible [329]. Distributed training introduces additional sources of variability (e.g., client sampling, communication rounds, local epochs); reproducibility therefore benefits from explicit reporting of protocols, aggregation schemes, and any privacy mechanisms applied [424, 175].

Within this dissertation, privacy-preserving and distributed learning is both an enabling technology and a design lens [175]. In Part I, federated learning supports collaboration across centers for stenosis-related tasks under data governance constraints, and the interaction between non-IID heterogeneity and model design is treated as central [198, 424, 73]. In Part III, privacy-aware sensing via radar provides an alternative route to protect identity while supporting sign language recognition in medical contexts [297]. Together, these perspectives reinforce a core claim of the thesis: privacy constraints are not only limitations to manage, but design drivers that motivate more deployable, inclusive, and institution-compatible AI systems through both learning-level and input-level strategies [175].

2.6 Evaluation, Calibration, and Clinical Considerations

In healthcare, an AI model is useful only when it supports reliable decisions under clinically meaningful operating conditions [348, 315]. This shifts emphasis away from single aggregated metrics and toward evaluation protocols that reflect real-world constraints: threshold-based decision rules, class imbalance and prevalence shift, calibration, external validation, and

robustness across sites and subpopulations [294, 64, 390, 348]. The goal is not merely to report performance, but to characterize when and why a system succeeds or fails, and whether its outputs can be trusted as part of a clinical pathway (screening, triage, monitoring, or decision support) [372, 213]. This section summarizes evaluation principles that recur throughout the dissertation, spanning classification, regression, calibration and uncertainty, and translation-oriented considerations [348, 315].

For classification tasks, including binary and multi-class settings (e.g., stenosis present/absent, severity categories, sign labels), conventional metrics capture different aspects of behavior [286, 348]. Accuracy is simple but can be misleading under imbalance, where predicting the majority class yields deceptively high values [137, 286]. Precision, recall, and the F1-score are more informative when false positives and false negatives have asymmetric consequences, and they encourage explicit reporting of error trade-offs rather than a single summary number [348, 286]. In clinical practice, sensitivity and specificity remain central because they correspond to true positive and true negative rates at a chosen threshold, aligning naturally with risk-driven and threshold-driven decision-making [348, 286]. Threshold-free summaries such as the area under the receiver operating characteristic curve (ROC-AUC) [131, 99] are widely used to summarize sensitivity-specificity trade-offs across operating points, while the area under the precision-recall curve (PR-AUC) [72, 318] is often more informative under class imbalance because it emphasizes the precision-recall regime relevant to rare findings. For detection and localization tasks, intersection over union (IoU) based criteria and mean average precision (mAP) [94] are standard, but clinical deployment typically also requires case-level analyses: whether lesions are missed in critical views, how errors distribute across anatomy, and whether failure modes correlate with acquisition quality or patient subgroups [348, 390, 372]. Across the dissertation, reporting is therefore treated as threshold-aware rather than metric-minimal: whenever clinically meaningful cut-offs exist, operating-point results (e.g., sensitivity/specificity at a defined threshold) are prioritized alongside global summaries [348].

For regression tasks involving continuous clinical indices (e.g., FFR/iFR estimation, cognitive score estimation), standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) quantify absolute deviations, while association measures such as Pearson correlation [284] and Spearman rank correlation [342] capture linear and monotonic relationships. The coefficient of determination (R^2) [86] provides an aggregate view of explained variance. However, many clinical endpoints derived from continuous indices are ultimately applied through thresholds (e.g., hemodynamic significance or screening cut-offs) [348]. For this reason, regression evaluation benefits from a dual interpretation: (i) continuous accuracy (error, correlation, bias) and (ii) thresholded decision performance

obtained by mapping predictions to clinically relevant categories [348]. Agreement analyses such as Bland-Altman plots [30] are often valuable when the focus is not only on correlation but on systematic bias and limits of agreement, especially when outputs may influence treatment decisions [348]. In this dissertation, this dual view is essential in cardiovascular functional assessment and in cognitive screening estimation, where continuous scores have direct clinical meaning but decisions are often threshold-driven [348].

Calibration is a critical complement to accuracy [123, 370]. A model can achieve strong average performance while being unsafe if it is poorly calibrated, i.e., if its predicted probabilities do not reflect empirical correctness [123, 370]. In clinical workflows, miscalibration can lead to inappropriate confidence: overconfident errors are particularly problematic under domain shift or when outputs are used to gate referral decisions [294, 123]. Calibration can be assessed using reliability diagrams, Expected Calibration Error (ECE) [123], and proper scoring rules such as the Brier score [117, 262]. When miscalibration is detected, post-hoc methods such as temperature scaling [123] (and more general probabilistic calibration such as Platt scaling) [291] can improve probability reliability without retraining, although calibration should be verified on representative data and, ideally, on external sites [64, 370]. In multi-center settings, calibration is not a single global property: it can vary by institution, device type, protocol, or prevalence, and it should therefore be measured with stratification whenever feasible [370, 64, 348].

Closely related is uncertainty-aware decision support [348, 191, 9]. In high-stakes settings, it can be preferable for a model to abstain or defer to a clinician when uncertainty is high rather than forcing a prediction [348]. Practical mechanisms range from simple confidence thresholding with referral to human review, to ensemble methods [191] and Monte Carlo dropout [110] that approximate predictive uncertainty, and conformal prediction approaches that provide prediction sets with formal coverage guarantees under stated assumptions [378, 9]. While not every contribution in this dissertation explicitly implements abstention, the evaluation philosophy is compatible with it: analyses are structured to expose uncertainty-linked failure modes (e.g., low-quality acquisitions, out-of-distribution patterns, rare classes) that would be natural triggers for referral strategies in future deployments [294, 348].

External validation and generalization are decisive for translation [64, 390, 348]. Domain shift in healthcare can arise from different devices and acquisition parameters, differences in demographics and disease prevalence, protocol variability across centers, and annotation practice differences [294, 390]. A model that performs well on internal splits may fail when deployed elsewhere if it has learned site-specific cues or if its calibration drifts under new prevalence [294, 370]. Mitigation strategies include cross-center evaluation with held-out sites, domain-aware training and harmonization, post-hoc recalibration on target cohorts

when appropriate, and careful reporting of inclusion criteria and data provenance to support reproducibility and auditability [64, 390, 348]. In addition, translation-oriented guidelines (e.g., TRIPOD/PROBAST for prediction models, and recent AI-specific reporting frameworks) encourage transparent description of datasets, splits, metrics, calibration analyses, and intended-use statements, which are necessary to interpret results within a clinical pathway [64, 390, 372, 213, 256].

Finally, clinical utility should be considered explicitly rather than inferred from aggregate performance [374, 348]. Decision-analytic tools such as decision curve analysis can relate model outputs to net benefit under plausible threshold probabilities and workflow assumptions, helping identify operating regions where a system is likely to improve care while avoiding harm [374]. Across this dissertation, these principles motivate evaluation designs that are aligned with clinical thresholds, emphasize reliability under heterogeneity, and prioritize evidence that supports safe deployment rather than purely benchmark-centric comparisons [348, 315].

Part I

AI for Cardiovascular Diagnostics

Chapter 3

Clinical Background and Challenges in Coronary Artery Assessment

Coronary artery disease (CAD) remains among the leading causes of morbidity and mortality worldwide. A central objective in CAD management is to identify coronary stenoses that are hemodynamically significant and therefore likely to benefit from revascularization. While severe stenoses are often evident, a substantial fraction of real-world cases falls into the category of intermediate lesions, where the relationship between anatomical narrowing and functional impairment is uncertain. In these scenarios, decision making based solely on visual inspection can be suboptimal, potentially resulting in unnecessary interventions or in missed ischemia-producing lesions.

In the catheterization laboratory, X-ray coronary angiography is the routine imaging modality used to visualize coronary anatomy and guide interventions. It provides high temporal resolution and direct visualization of contrast flow through the vessel lumen, making it the de facto reference for procedural navigation. At the same time, angiography is inherently a 2D projection of a 3D structure and is therefore affected by overlap, foreshortening, and view dependency. Image appearance and contrast dynamics further depend on acquisition settings, injection protocols, patient motion, and catheter positioning. These sources of variability complicate consistent anatomical quantification, especially for borderline stenoses and complex anatomies.

Because anatomical narrowing does not uniquely determine downstream ischemia, physiological indices have become central for functional assessment. FFR [288] and iFR [325] are widely used measures to quantify the functional impact of a coronary lesion. They are typically obtained invasively using a pressure wire and, in the case of FFR, pharmacological hyperemia. These measurements provide a more direct link to clinical outcomes and

support evidence-based decisions regarding percutaneous coronary intervention. However, invasive physiological assessment introduces additional procedural steps, costs, and practical barriers. In routine practice, this can limit adoption when workflow time is constrained, when equipment is not readily available, or when the perceived benefit is uncertain. As a consequence, a persistent gap remains between guideline-supported functional assessment and broad real-world utilization, particularly for intermediate stenoses.

A non-invasive or minimally invasive estimation of functional significance directly from plain angiography has the potential to reduce acquisition barriers while preserving the clinical objective. If functional assessment could be inferred from routinely acquired angiographic sequences, it could enable broader and more consistent use of physiology-informed decision making without additional instrumentation. Within the scope of this dissertation, this motivates a key translational question: whether angiographic videos contain sufficient spatio-temporal information to estimate not only anatomical stenosis severity, but also the functional consequences of lesions. Contrast propagation, vessel opacification dynamics, and subtle temporal patterns may encode physiological correlates that are not fully captured by static, frame-level representations.

Developing AI systems for stenosis and functional assessment from angiography entails a combination of clinical, data, and modeling challenges. First, intermediate lesions embody clinical ambiguity: small differences in lumen narrowing can correspond to large differences in functional impact depending on lesion length, vessel size, microvascular status, collateral circulation, and downstream demand, making the mapping from image appearance to physiology intrinsically complex and patient-dependent. Second, view dependency and 2D projection effects remain limiting factors: angiography is sensitive to projection angle, overlap with other vessels, and foreshortening, and even sequences of the same lesion can appear substantially different across acquisitions; when only a single view is available, it may provide an incomplete representation of the underlying 3D geometry. Third, angiographic sequences exhibit temporal variability and procedural artifacts, including motion from cardiac contraction and respiration, variable contrast injection timing and rate, table or C-arm motion, and transient occlusions or catheter-induced artifacts, which complicate temporal alignment and can confound purely appearance-based models.

Additional constraints arise from supervision and data. Functional labels such as FFR/iFR are available only when invasive measurements are performed and may reflect measurement variability, operator choices, and patient-specific hemodynamics. Labels are typically assigned at the lesion or vessel level rather than per frame, yielding weak supervision for video models. Anatomical stenosis severity, when used as an auxiliary target, is often derived from visual estimation and quantitative coronary angiography, both of which exhibit inter-

observer variability. Moreover, multi-center datasets introduce heterogeneity across acquisition protocols, imaging systems, patient populations, and clinical practices, creating domain shift that can degrade generalization unless addressed explicitly through robust training and evaluation strategies. Finally, any system intended for procedural decision support must satisfy cath-lab deployment constraints: timely inference, stable performance under variability, meaningful confidence and calibration, and interfaces that support clinical interpretation and oversight.

Part I addresses these challenges through a progression of contributions. It develops spatio-temporal deep models that exploit the dynamics of angiographic sequences for stenosis assessment and functional estimation, investigates privacy-preserving and distributed learning strategies to enable multi-center collaboration without centralizing sensitive data, and studies foundation-model-based approaches and prompt learning to improve generalization under limited supervision and domain shift. These methodological advances are then framed within an end-to-end system perspective, including multi-center evaluation and the associated intellectual property activities related to angiographic image processing.

Chapter 4

Related Work on Coronary Stenosis Evaluation

Research on coronary stenosis assessment from imaging spans a continuum from anatomical detection and grading to functional evaluation through physiology-inspired indices such as FFR and iFR. In parallel, real-world adoption introduces constraints that are central in healthcare AI, including data fragmentation across institutions, privacy requirements, and robustness under domain shift. This chapter reviews prior work along four axes that structure Part I of this dissertation: (i) frame-based and detection-oriented analysis of coronary angiography, (ii) video-level spatio-temporal modeling and hybrid architectures, (iii) privacy-preserving multi-center learning, and (iv) foundation models and parameter-efficient adaptation for generalization.

A large portion of the literature follows 2D pipelines that analyze individual frames extracted from angiography videos and aggregate predictions through late fusion or voting. Many methods perform stenosis grading into a small number of severity levels, or classify lesions as hemodynamically significant by thresholding FFR/iFR values. Key-frame selection is commonly addressed with CNN-based approaches [258, 316, 10] or convolutional-recurrent designs that incorporate limited temporal context while still producing frame-centric representations [66, 224, 276, 65]. To reduce spurious correlations and constrain the region of interest, several pipelines include vessel segmentation or coronary-tree extraction so that learning is biased toward the lumen and contrast-filled regions rather than surrounding anatomy [393, 16].

Closely related efforts target stenosis detection and localization on angiographic frames. These approaches typically first identify a key frame and then apply object detection models to localize stenoses [393, 70, 87, 281], with comparative evaluations of detection architectures reported in [69]. Other strategies emphasize vessel shape and appearance analysis, often

coupling segmentation with geometric descriptors and downstream classification [418, 417, 380]. Interpretability tools are also employed to produce activation maps that support plausibility checks, guide detection, or justify frame-based decisions in coronary imaging [258, 66, 65]. Overall, 2D pipelines have benefited from simpler training setups and the ability to reuse mature image backbones, but their reliance on key-frame selection and limited modeling of contrast-flow dynamics motivates video-level approaches.

The move from anatomical severity toward functional significance is driven by the clinical reality that luminal narrowing does not uniquely determine downstream ischemia, particularly in intermediate lesions. Invasive measurements of FFR and iFR provide guideline-supported tools for physiology-informed decisions [183, 266], and clinical studies have reported benefits when revascularization is guided by physiology rather than anatomy alone [426, 359, 101, 71, 38, 118]. However, adoption can be limited by procedural overhead and cost, motivating non-invasive estimation strategies [282]. Historically, many non-invasive approaches have focused on CT-derived pipelines and computational modeling, which can require extensive preprocessing and coronary reconstruction [259, 156]. Angiography-only learning is appealing because it reuses routine acquisition and, in principle, can integrate into cath lab workflows without additional instrumentation.

Recent work has begun to explore learning-based quantitative coronary analysis (QCA) and related regression targets from angiography sequences. In particular, [410, 409] regress multiple clinical indices using a primary angiography view plus an additional side view and a manually selected key frame, employing shared 3D convolutional backbones for multi-view processing and attention-based fusion between video features and key-frame features [409]. Related efforts consider stenosis detection on full angiography sequences [396, 129, 130]. These studies provide evidence that angiographic dynamics can support richer estimation beyond purely anatomical grading, but they often introduce additional acquisition assumptions (e.g., multiple views) or manual steps (e.g., key-frame selection). This motivates models that operate directly on routine single-view sequences with minimal operator burden and that can be trained under the weak supervision typical of clinical practice.

Spatio-temporal learning is central to these video-level approaches. Compared to frame-based pipelines, 2D+t methods model the full angiography sequence, aiming to represent contrast propagation and temporal patterns that may correlate with functional significance [410, 409, 396, 130]. 3D CNNs provide an inductive bias for local spatio-temporal feature extraction and have been used broadly for medical video understanding. Attention mechanisms can then be employed to emphasize salient spatial regions and critical temporal phases, potentially improving stability under projection changes and background clutter. More broadly, there is a growing trend toward hybrid CNN-attention architectures that

combine convolutional hierarchies with attention blocks. Non-local operators introduced attention-style interactions within convolutional networks [383]. Subsequent designs integrate convolution into transformer-like processing through convolutional token embeddings and projections, as in CvT [391], or through hierarchical representations induced by windowing and patch merging, as in Swin Transformers [215, 214]. In medical imaging, hybrid designs are often motivated by the need to preserve local texture sensitivity while retaining global context and interpretability properties associated with attention-based reasoning [264, 54, 344]. In the specific context of coronary angiography, these insights motivate architectures that can jointly represent local lesion appearance and global flow dynamics while remaining compatible with real-world acquisition constraints.

Practical translation also depends on learning from multi-center data under governance constraints. Medical data are inherently distributed and heterogeneous, with non-IID patient populations, protocol variation, and site-specific artifacts. FL [231] addresses this setting by enabling collaborative training without centralizing raw data. In the standard paradigm, client nodes compute local updates and a server aggregates them iteratively. FedAvg [231] performs simple averaging, FedProx [198] adds a proximal term to reduce client drift, robust aggregation strategies such as Federated Median reduce sensitivity to outliers [289], and FedBN keeps batch normalization parameters local to mitigate domain shift across clients [202]. FL has been explored for detection and segmentation, often adapting architectures such as YOLOv3 [308] and Faster R-CNN [311] to federated settings [159, 221], or applying FedAvg-style training to SSD [211] with additional mechanisms to address non-IID effects [404]. In medical segmentation, federated variants of U-Net [317] and strategies for weight sharing and disentanglement have been studied to handle imbalance and domain shift across hospitals [401, 356, 199, 25]. Because multi-center constraints are pervasive, healthcare has become a practical benchmark domain for FL methods [102, 73, 199]. Within angiography, FL is particularly relevant when generalization requires exposure to diverse acquisition protocols and cohorts, but data cannot be centralized due to regulatory and institutional constraints.

In parallel to FL, foundation models and parameter-efficient adaptation offer a complementary route to improve generalization under domain shift. Large-scale contrastive vision-language pretraining has produced models that align images and text in a shared representation space, including CLIP [295] and ALIGN [162], and subsequent architectures such as BLIP [197], Florence [405], CoCa [403], and SimVLM [389]. These systems commonly rely on transformer backbones [373] and web-scale paired data. However, domain shift is substantial in medical imaging, where fine-grained visual patterns and specialized terminology are underrepresented in generic pretraining corpora. Domain-specific vision-language models

and clinical pretraining efforts attempt to reduce this gap [32, 416, 399, 413, 387, 17, 233], yet direct fine-tuning can be compute-intensive and prone to overfitting under limited labeled data.

These constraints have made parameter-efficient adaptation central. Adapter layers [147], BitFit [407], and LoRA [148] update only small parameter subsets. Prompt-based methods further reduce trainable degrees of freedom by optimizing learnable context tokens. Visual Prompt Tuning (VPT) [163] introduces learnable visual tokens, while CoOp [423] and CoCoOp [421] optimize textual prompts, with instance-adaptive variants improving robustness. Prompt learning also has deep roots in NLP, from template search [166] and discrete token optimization (AutoPrompt) [333] to continuous approaches such as prefix tuning [200], prompt tuning [194], P-tuning v2 [212], and factual probing [419]. At the same time, prompt design can be brittle, and small wording changes can lead to large performance differences in medical tasks [415]. These observations motivate approaches that learn prompts, and when appropriate learnable tokens on both modalities, to adapt vision-language priors to angiography and stenosis-related objectives with limited supervision and improved robustness.

A recurring limitation across prior angiography-based studies concerns the scale, composition, and acquisition conditions of the datasets used for evaluation. Many methods are developed on relatively small cohorts, often collected within a single center or under limited protocol variability, which can restrict the assessment of robustness across scanners, acquisition settings, operators, patient populations, and lesion distributions. In addition, several pipelines rely on controlled assumptions such as manually selected key frames, predefined views, or curated image quality, which simplify the learning problem but may overestimate performance in routine cath-lab conditions. These dataset-related constraints are particularly relevant for coronary angiography because domain shift can arise not only from imaging hardware, but also from contrast injection protocols, projection angle, frame rate, motion artifacts, and local clinical practice. As a consequence, high performance on internal or narrowly curated datasets does not necessarily imply reliable generalization to external centers or heterogeneous acquisition workflows. This limitation motivates the emphasis of Part I on video-level modeling, reduced dependence on manual frame selection, multi-center evaluation, and privacy-preserving collaborative learning as complementary strategies to improve generalization under real-world data variability.

Overall, prior work establishes strong baselines for frame-based stenosis analysis and provides emerging evidence that spatio-temporal learning on angiography sequences can support richer anatomical and functional assessment. Clinical translation, however, requires robustness across sites and protocols, making privacy-preserving collaboration via federated learning practically relevant. Foundation-model representations and parameter-efficient adap-

tation provide a further path to strengthen generalization under domain shift, particularly when prompt learning is used to align medical semantics with limited labeled supervision. The methods in Part I build on these directions by modeling angiography as a spatio-temporal signal for functional assessment, addressing multi-center constraints through federated learning, and leveraging foundation models and prompt-based adaptation to improve transfer and deployment robustness.

Chapter 5

Coronary Artery Stenosis Assessment in X-Ray Angiography for Non-Invasive Evaluation

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Federica Proietto Salanitri, Giovanni Bellitto, Ovidio De Filippo, Fabrizio D'Ascenzo, Simone Palazzo, and Concetto Spampinato. Coronary artery stenosis assessment in x-ray angiography through spatio-temporal attention for non-invasive ffr and ifr estimation. In *BIOSTEC (1)*, pages 305–312, 2024
- Ovidio De Filippo, Raffaele Mineo, Michele Millesimo, Wojciech Wańha, Federica Proietto Salanitri, Antonio Greco, Antonio Maria Leone, Luca Franchin, Simone Palazzo, Giorgio Quadri, et al. Non-invasive physiological assessment of intermediate coronary stenoses from plain angiography through artificial intelligence: the starflow system. *European Heart Journal-Quality of Care and Clinical Outcomes*, 11(3):343–352, 2025

5.1 Overview

Physiology-guided assessment of intermediate-grade coronary stenoses is recommended to support decision-making for percutaneous coronary intervention (PCI), with invasive FFR and iFR measurements used to determine whether a lesion is functionally significant [267, 183]. Although FFR/iFR-guided PCI has shown clinical benefit and can reduce unnecessary revascularization, these techniques remain underused in routine practice, largely due to the time required for setup and measurement, the cost of pressure-wire equipment, workflow

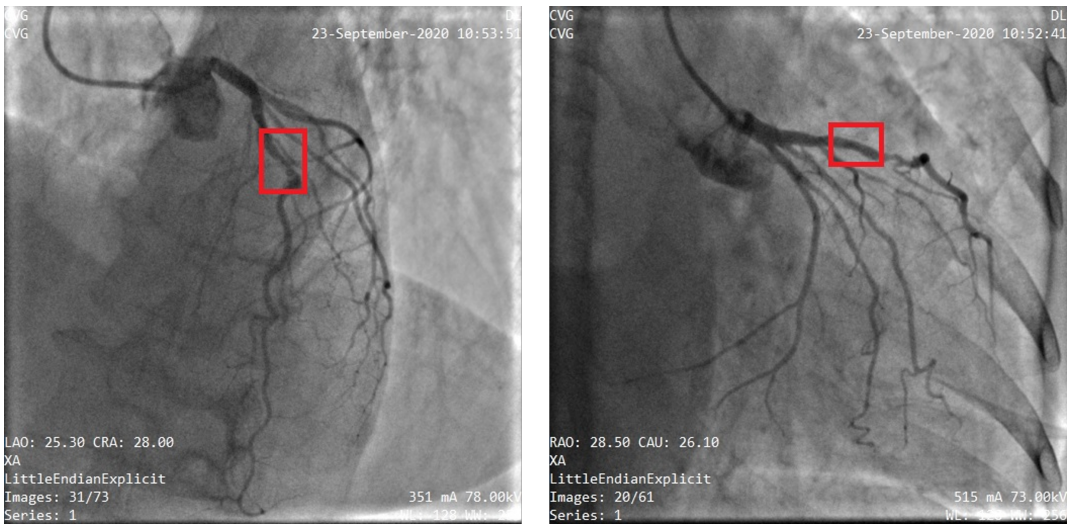


Figure 5.1: Two sample angiography views for the same patient. Red bounding boxes show the major stenosis.

constraints in busy catheterization laboratories, and the non-zero procedural risk of invasive assessment. In addition, clinical decisions are often driven by whether FFR/iFR fall below established cutoffs (0.80 for FFR and 0.89 for iFR) rather than by their exact absolute values [267, 359, 75].

These barriers motivate approaches that approximate physiological assessment while reusing routinely acquired data. Coronary angiography is performed in essentially all patients undergoing invasive coronary evaluation and contains dynamic information about contrast transport through the coronary tree. If functional significance could be inferred directly from plain angiographic sequences, physiological assessment could become more accessible and easier to integrate into standard workflows. However, existing automated strategies often rely on strong operator involvement and/or non-trivial preprocessing. Many deep-learning approaches are *frame-centric* (2D), typically requiring key-frame selection and then applying late fusion or voting, while more recent spatio-temporal models attempt to extract features directly from the full video. In clinical-grade functional estimation, influential prior pipelines have also required multiple angiographic projections and, in some cases, manually chosen key frames [410, 409], increasing the burden on both clinicians and patients (Fig. 5.1).

We study whether artificial intelligence can non-invasively estimate lesion functional significance, both as continuous indices and as dichotomous decisions at standard clinical thresholds, directly from plain coronary angiography. Concretely, we develop a multi-task deep model that processes angiographic videos acquired from two projections of the coronary of interest, avoids manual key-frame selection, and jointly predicts FFR and iFR when available. The architecture combines a 3D convolutional backbone to model local spatio-

temporal patterns of contrast flow with self-attention layers to capture broader dependencies across the extracted feature set. Training is supervised using both FFR and iFR; missing targets are handled through multi-branch outputs, enabling learning from heterogeneous acquisition protocols and partially observed labels. In addition to functional estimation, the system provides lesion-level classification (clinically significant vs non-significant) and includes an explainability component to highlight the temporal and spatial regions of the angiographic sequence that contribute most to the prediction.

The approach was validated on a multi-center Italian cohort comprising 778 angiography exams from 389 consecutive patients enrolled across five centers. The STARFLOW clinical study reports high performance for both continuous estimation and threshold-based decision support: overall accuracy of 87.3% (95% CI 81.2%-93.4%), sensitivity 82.4% (71.9%-96.4%), and specificity 92.2% (90.4%-93.9%). Subgroup analyses show accuracy of 84.8% (77.8%-91.8%) for FFR and 90.2% (86.0%-94.6%) for iFR, with corresponding sensitivity and specificity reported in the clinical paper. These findings support the feasibility of deriving physiology-informed decision cues from routinely acquired angiography without invasive probes, and suggest a pathway toward workflow-compatible decision support for intermediate lesions.

Our contributions are (i) an end-to-end spatio-temporal deep model for angiography-based functional assessment that does not require manual key-frame selection, (ii) a multi-branch multi-task formulation enabling joint learning of FFR/iFR and dichotomous classification under partially missing targets, and (iii) a multi-center clinical validation (STARFLOW) demonstrating strong accuracy, sensitivity, and specificity in an unselected cohort, complemented by explainability analyses to improve transparency of model behavior.

5.2 Method

5.2.1 Patient cohort, reference standard, and labeling

Consecutive patients undergoing invasive physiological assessment of intermediate coronary stenoses (angiographic visual estimation of diameter stenosis $\geq$30% and $\leq$90%) with iFR, FFR, or both were retrospectively included from January 2020 to January 2022 across five centers. Both Chronic Coronary Syndrome (CCS) and Acute Coronary Syndrome (ACS) presentations were considered; in ACS patients, physiology was assessed only on non-culprit lesions. Coronary angiography and wire-based measurements were performed according to standard practice at each site.

Physiological assessment followed the conventional pressure-wire protocol (Philips

Volcano®). The wire was equalized at the guide catheter, advanced at least three vessel diameters distal to the stenosis, and distal and aortic pressures were recorded. iFR was computed through an automated algorithm applied to time-aligned pressure traces over the diastolic wave-free period across a minimum of five beats, before adenosine administration [325]. Maximal hyperaemia was induced via continuous intravenous adenosine (140 μg/kg/min for 2 min) or intracoronary bolus injections (100-200 μg), and FFR was computed as the ratio of mean distal coronary pressure to mean aortic pressure during stable hyperaemia (lowest average of three consecutive beats). For each case, expert clinicians labeled the coronary segment undergoing invasive testing and identified the corresponding stenosis on angiography. Patients provided informed consent for anonymized data use; the study was approved by the institutional review boards of the leading and participating centers, where required.

5.2.2 Angiography selection and preprocessing

For each patient, two angiographic projections of the coronary of interest were selected by trained cardiologists. Raw frame sizes ranged from 512×512 to 1024×1024 pixels, frame rates from 15 to 30 fps, and clip duration from 3 to 5 s, reflecting multi-center acquisition heterogeneity. To obtain a standardized input suitable for robust training, all angiographies were (i) resampled to 15 fps, (ii) resized to 256×256 pixels, and (iii) temporally clipped to a fixed 4 s duration (60 frames). This normalization reduces spurious variability due to acquisition settings and enforces a consistent temporal scale for modeling contrast-flow dynamics. In deployment, the same steps can be executed automatically as part of the ingestion pipeline.

5.2.3 Model architecture

The proposed model (Fig. 5.2) combines a shared spatio-temporal convolutional backbone with an explicit space-time attention module and a multi-task output design. Each of the two views is processed by the same 3D convolutional feature extractor inspired by 3D ResNet-18 [364]. The backbone is initialized from Kinetics-400 pretraining [48]. To adapt from RGB videos to X-ray angiography, the first-layer convolutional kernels are averaged across the channel dimension, enabling effective processing of single-channel inputs.

For each view, the backbone outputs a spatio-temporal tensor of shape $[W, H, T, C]$, where W, H, T, and C denote width, height, time, and channels. The tensor is serialized into $[W \times H \times T, C]$ tokens and refined by a multi-head spatio-temporal self-attention module,

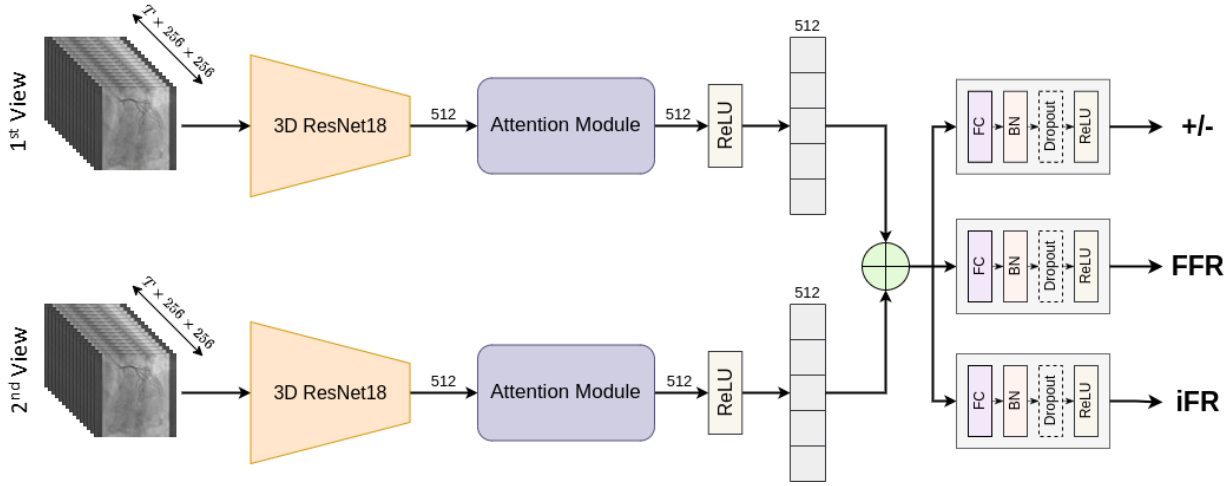


Figure 5.2: Architecture overview. A 3D CNN, common to all views, extracts spatio-temporal features which are subsequently refined through a multi-head spatio-temporal self-attention. The resulting features are then channeled to a three-output layer which carries out binary classification (over the threshold), and quantification of iFR and FFR.

implemented as a compact space-time transformer block:

$$\mathrm{MHA}(Q, K, V) = \left\|\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h\right\| \times W_O \tag{5.1}$$

with

$$\mathrm{head}_i = \mathrm{SelfAttention}(QW_{Qi}, KW_{Ki}, VW_{Vi}) \tag{5.2}$$

and

$$\mathrm{SelfAttention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.3}$$

where d_k is the key dimension and h is the number of heads. This attention stage promotes long-range space-time interactions, allowing the model to emphasize diagnostically relevant contrast-flow patterns while suppressing background and acquisition artifacts.

The attended representation is fed to a three-branch output layer that jointly supports: (i) binary classification of hemodynamic significance and (ii) regression of continuous FFR and iFR values. Binary labels are defined using established thresholds (FFR 0.80 and iFR 0.89) [267, 359, 75, 23]. Since not all samples provide both continuous targets, the regression heads are optimized conditionally: the FFR head is updated only when wire-based FFR is available, and analogously for iFR. This masked multi-task formulation enables training on heterogeneous clinical records without discarding partially labeled cases.

5.2.4 Training objective

Training minimizes a weighted combination of binary cross-entropy for classification and L1 losses for regression:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{5.4}$$

$$L_{L1} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{5.5}$$

$$L_{\text{total}} = \alpha L_{BCE} + \beta L_{L1} \tag{5.6}$$

where α and β weight the two objectives. In practice, L_{L1} is computed only over the regression head(s) whose target is available for the current sample, ensuring that missing labels do not bias optimization.

5.2.5 Implementation and optimization protocol

Input angiographies are normalized to $[0, 1]$ and then standardized to zero mean and unit variance. Data augmentation consists of random horizontal and vertical flips and random 90-degree rotations, applied identically to all frames within a clip to preserve temporal coherence. Training is performed with AdamW (learning rate 1e-5), batch size 8, for 300 epochs, using automatic mixed precision. Experiments are implemented in PyTorch with MONAI support and executed on two NVIDIA Tesla T4 GPUs.

5.2.6 Clinical covariates extension and optional late fusion

In addition to purely visual inference, we evaluate whether a limited set of non-identifying clinical variables can improve classification performance. When enabled, clinical features are encoded as follows: age and left ventricular ejection fraction (LVEF) are discretized into 20 bins and represented via one-hot vectors; the indication for coronary angiography (STEMI, NSTEMI, unstable angina, stable angina, positive stress test) is represented as a 5-dimensional vector (1 if present, -1 otherwise). Additional covariates (e.g., chronic kidney disease, ACS vs CCS) are included as binary indicators when available. Clinical inputs are embedded through a learnable fully connected layer into a 256-dimensional vector and fused late with the visual representation before the output heads, preserving the same multi-branch decision structure.

5.2.7 Evaluation protocol and endpoints

Model performance is assessed through five-fold nested cross-validation. For each split, data are partitioned into 60% training, 20% validation, and 20% test, maintaining the original label proportion. The primary classification endpoint is balanced accuracy, computed as (sensitivity+specificity)/2, alongside sensitivity and specificity. Discrimination is further summarized using ROC-AUC and precision-recall curves. For continuous estimation, regression quality is reported via mean squared error (MSE) and mean absolute error (MAE). Robustness is additionally examined through sensitivity analyses across centers and X-ray systems, and by comparison with relevant two-view approaches for angiography-based physiological assessment [410, 409, 396].

5.3 Experimental Results

5.3.1 Dataset and cohort construction

The proposed approach was developed and evaluated on a multi-center cohort collected between January 2020 and January 2022. A total of 763 consecutive patients undergoing invasive physiological assessment for intermediate coronary stenoses were screened; 389 patients were retained after a data quality check, which excluded: (i) cases with only one angiographic view available (STARFLOW requires two projections), (ii) cases with insufficiently reliable invasive iFR/FFR traces, and (iii) cases for which only rendered video was available (missing raw data) (Fig. 5.3). For each retained patient, two angiographic projections of the coronary of interest were selected, yielding 778 angiographic video sequences overall. The IRB protocol number is 0092163.

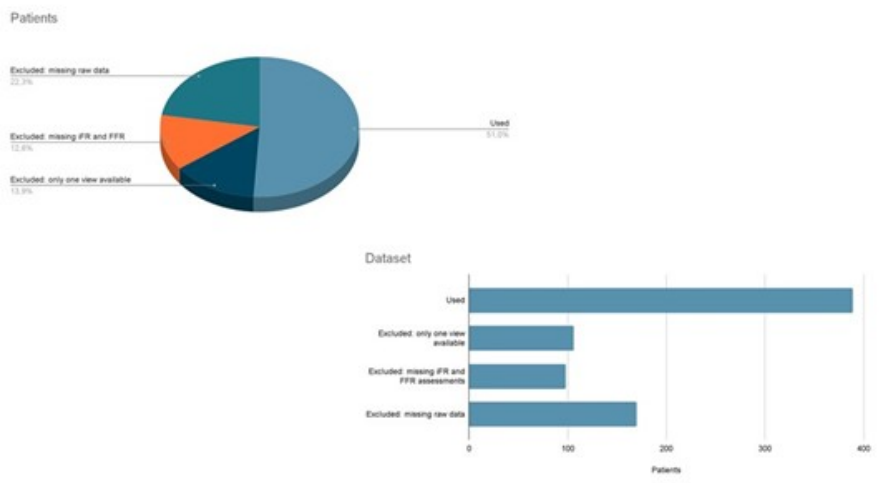


Figure 5.3: Patients excluded from the dataset used.

Across the final cohort ($n = 389$), mean age was 67.9 ± 9.6 years and 86 patients (22%) were female. Both CCS and ACS cases were included; in ACS, physiology was assessed only on non-culprit lesions. FFR values were available for 251 patients (64.5%), iFR values for 228 patients (58.6%), and both indices for 90 patients (23.1%). Using standard thresholds (FFR < 0.80 and iFR < 0.89), 93 patients (23.9%) were labeled as haemodynamically significant, resulting in a marked class imbalance. Angiographies were acquired with heterogeneous protocols (spatial resolution 512×512 to 1024×1024, frame rate 15 to 30 fps) and were standardized prior to learning (resized to 256×256, resampled to 15 fps, clipped to 60 frames/4 seconds).

5.3.2 Baseline characteristics

Baseline and procedural characteristics are summarized in Table 5.1. The majority of lesions were located in the left anterior descending artery (LAD, 71.7%). Mean diameter stenosis was $59.8 \pm 11.1\%$, with its distribution shown in Fig. 5.4. Mean invasive FFR and iFR were 0.84 ± 0.08 and 0.90 ± 0.07, respectively.

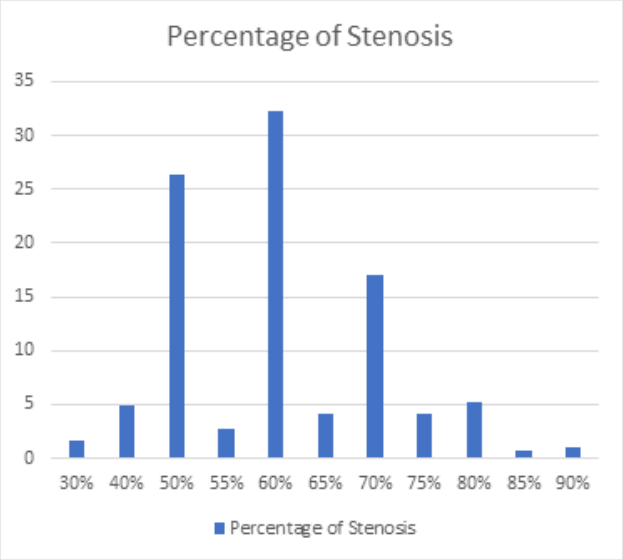


Figure 5.4: Distribution of percentage of stenosis across the final cohort.

5.3.3 Evaluation protocol and metrics

All models were evaluated through 5-fold nested cross-validation (60% training, 20% validation, 20% test per split), preserving the original label proportion. Given the class imbalance, performance is primarily reported via balanced accuracy, together with ROC-AUC, sensitivity, and specificity. For regression of continuous indices, MSE and MAE are used:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.8}$$

where y_i and $\hat{y}_i$ denote reference and predicted values.

5.3.4 Classification performance and comparison with prior methods

Table 5.2 reports the comparison against general deep video architectures adapted to the task (S3D [395], MVCNN [350], GVCNN [103]) and clinical AI methods designed for angiography-based quantification (DMQCA [410], HEAL [409], DMTRL [396]). The proposed method achieves a balanced accuracy of 87.3, improving over the closest comparator (81.7). ROC

Table 5.1: Baseline features: STEMI, ST-segment elevation myocardial infarction; NSTEMI, non ST-segment elevation myocardial infarction.

Baseline features	**n = 389**
Age, (years)	67.9 ± 9.61
Female	86 (22%)
Chronic kidney disease	66 (17%)
Left ventricular ejection fraction, %	52 ± 10
Indication for coronarography	
STEMI	23 (5.9%)
NSTEMI	62 (16.1%)
Unstable angina	67 (17.2%)
Stable angina	125 (32.1%)
Positive stress test	112 (28.7%)
Vessel	
Left main	13 (3.3%)
Left anterior descending	279 (71.7%)
Left circumflex	44 (11.2%)
Ramus intermedius	10 (2.5%)
Diagonal branch	8 (2.2%)
Obtuse marginal branch	19 (4.9%)
Postero-lateral branch	3 (0.8%)
Right coronary artery	13 (3.4%)
Lesion location	
Proximal	52.2%
Mid	43.4%
Distal	4.4%
Length of lesion, mm	18.4 ± 14.5
Diameter of stenosis, %	59.8 ± 11.1
Fractional flow reserve	0.84 ± 0.08
Instantaneous wave-free ratio	0.90 ± 0.07
Adenosine administration route[1]	
Intracoronary	227 (90.4%)
Intravenous	24 (9.6%)

and precision-recall curves against clinical AI baselines are shown in Fig. 5.5, confirming improved discrimination under the same operating conditions.

Table 5.2: Comparison of our method with state-of-the-art general deep learning and clinical AI methods

Methods	Accuracy	AUC	Sensitivity	Specificity
S3D [395]	79.5±5.54	0.93±0.03	65.4±14.17	93.6±4.81
MVCNN [350]	84.5±6.11	0.85±0.07	76.8±10.20	92.2±2.18
GVCNN [103]	78.4±4.52	0.87±0.05	71.0±7.29	85.8±4.36
DMQCA [410]	81.7±3.69	0.82±0.04	70.2±8.00	93.2±1.87
HEAL [409]	79.5±5.18	0.84±0.05	67.9±9.32	91.4±3.26
DMTRL [396]	79.1±3.79	0.85±0.05	67.0±6.56	91.2±3.48
Ours	87.3±6.14	0.93±0.03	82.4±12.24	92.2±1.79

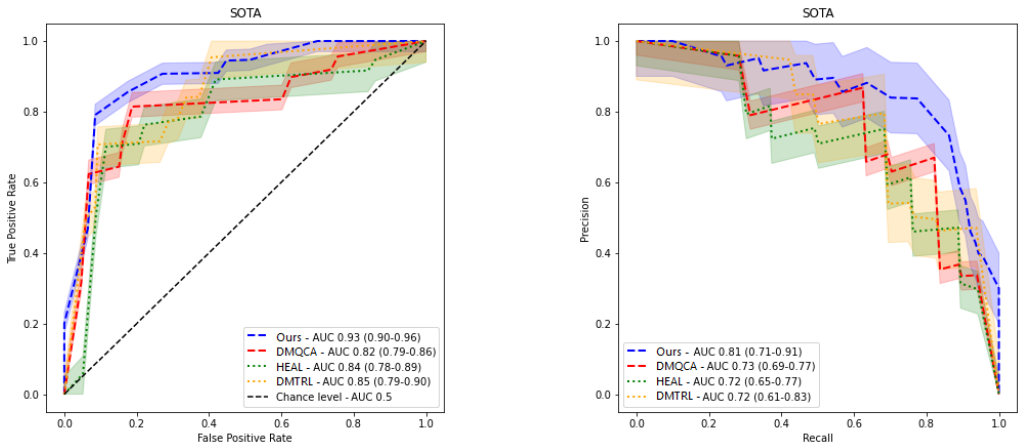


Figure 5.5: ROC (left) and precision-recall (right) curves comparing our approach with state-of-the-art clinical AI methods.

To report clinically oriented operating characteristics, Table 5.3 summarizes balanced accuracy with 95% confidence intervals, together with PR score, PPV, and NPV, and provides stratified results for FFR-only and iFR-only subgroups. Over the full cohort, balanced accuracy is 87.3% (81.2-93.4%), sensitivity 82.4% (71.9-96.4%), and specificity 92.2% (90.4-93.9%), with ROC-AUC 0.93 (0.90-0.96). Performance is higher in the iFR subgroup (balanced accuracy 90.2%, ROC-AUC 0.98) than in the FFR subgroup (balanced accuracy 84.8%, ROC-AUC 0.79), reflecting the different distributions and uncertainty profiles of the two indices.

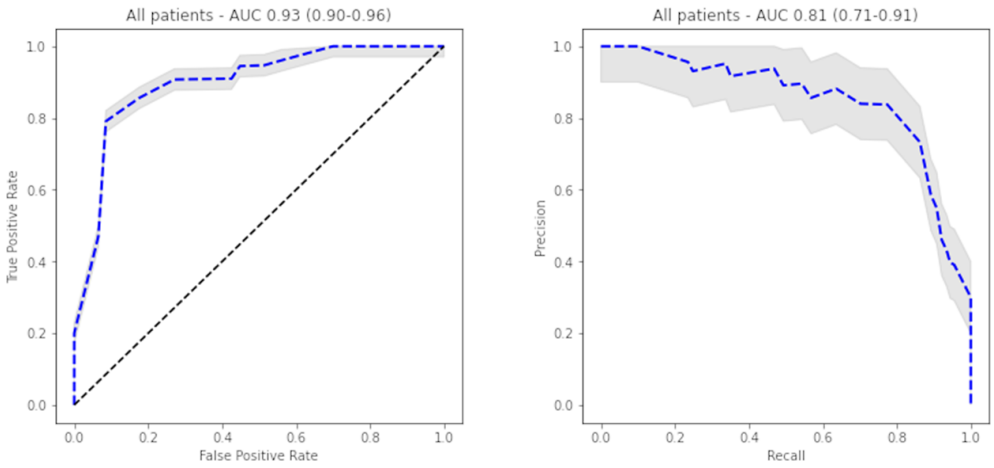


Figure 5.6: AUC-ROC (left) and precision-recall curve on a fold (right) for all patients.

Table 5.3: Performance of STARFLOW for overall patients, FFR patients, and iFR patients.

Methods	**Acc.**	**AUC**	**PR**	**Sens.**	**Spec.**	**PPV**	**NPV**
STARFLOW	87.3±6.10	0.93±0.03	0.91±0.04	82.4±12.25	92.2±1.75	89.2±4.50	87.5±5.40
STARFLOW FFR	84.8±7.00	0.79±0.13	0.79±0.16	81.9±12.50	87.7±2.20	63.7±11.10	94.7±2.20
STARFLOW iFR	90.2±4.30	0.98±0.02	0.98±0.02	87.2±10.60	93.2±1.50	87.2±7.00	93.2±1.60
STARFLOW + Cl. Data	84.3±6.30	0.91±0.04	0.88±0.05	69.9±12.95	98.7±1.30	97.7±3.40	84.2±4.80

5.3.5 Regression performance and runtime

When evaluated as a regression problem, the model attains low errors for both indices (Table 5.4), supporting the consistency between continuous estimation and the dichotomous decision rule. Inference is efficient: the model processes a 60-frame (4 s) two-view exam in approximately 62.56 ms in the reported setup, indicating feasibility for near-real-time clinical decision support.

Table 5.4: Performance in terms of MSE and MAE when regressing FFR and iFR values

Measure	MSE	MAE
FFR	0.060±0.005	0.037±0.002
iFR	0.045±0.005	0.026±0.003

5.3.6 Impact of clinical covariates

We assessed late fusion of a small set of clinical covariates (age, LVEF, chronic kidney disease, ACS vs CCS) with the visual representation. As reported in Table 5.3, adding clinical data

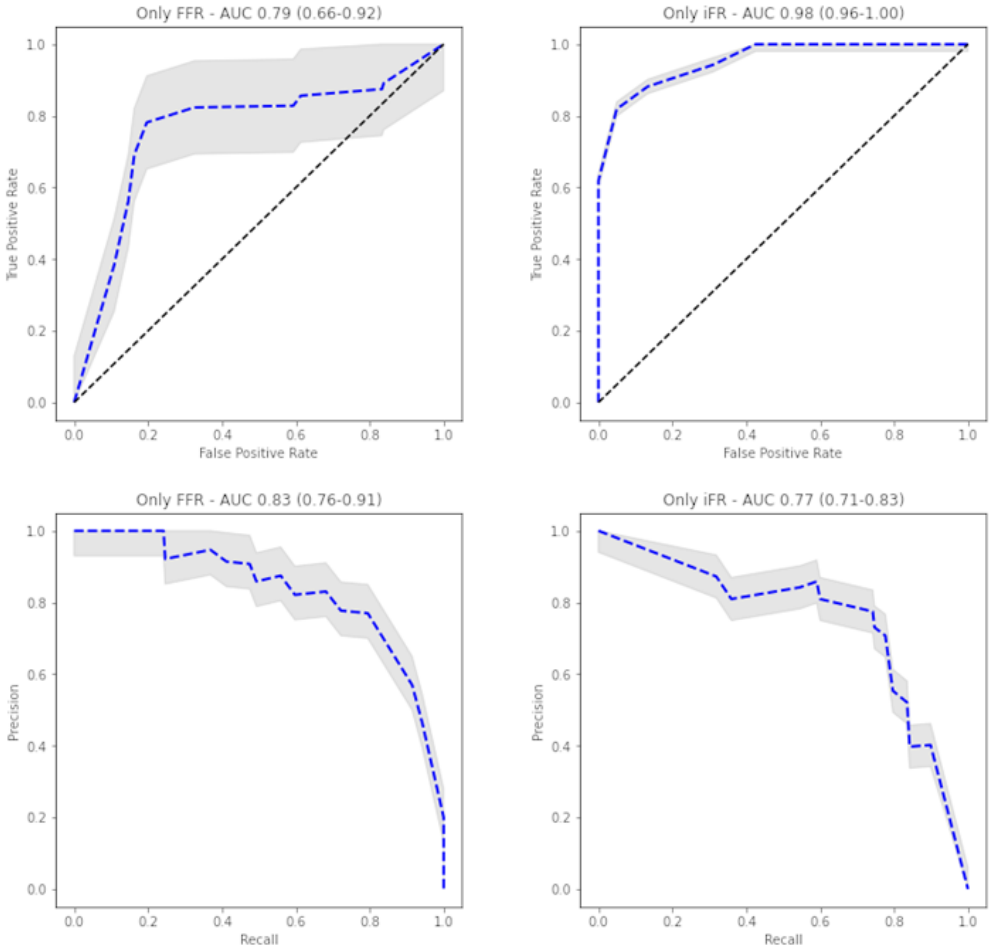


Figure 5.7: AUC-ROC (above) and precision-recall curve on a fold (below) for FFR patients (left) and for iFR patients (right).

shifts the operating point toward higher specificity (98.7%) at the cost of reduced sensitivity (69.9%), and decreases balanced accuracy relative to the visual-only model. This suggests that, in the present setting, the hybrid representation tends to degrade the visual latent space used for physiology-oriented discrimination, and that careful feature selection and fusion design are needed when incorporating structured covariates.

This result should not be interpreted as evidence that clinical variables are intrinsically irrelevant for physiology estimation. Rather, it highlights the limitations of a simple late-fusion strategy under limited data. A possible explanation is representation imbalance: high-dimensional angiographic embeddings and compact tabular variables may contribute at different scales, so simple concatenation can either make the model ignore the clinical branch or allow noisy covariates to shift the decision boundary in an unstable way. This effect is plausible in the present cohort, where the number of samples is limited and the fused model introduces additional degrees of freedom. The reduction in sensitivity also suggests that the

clinical branch may have shifted the operating point toward a more conservative behavior, improving specificity but increasing false negatives. More structured fusion strategies, such as modality-specific normalization, bottleneck projections, gated fusion, FiLM-style conditioning, cross-attention with clinical tokens, or calibrated decision-level fusion, may better exploit clinical context while limiting overfitting.

5.3.7 Comparison with multi-view clinical baselines and robustness across acquisition systems

Figure 5.8 and Table 5.5 provide an additional view of the comparison against multi-view methods for angiography-based physiological assessment, namely Direct Quantification [410], Direct Quantification with Attention [409], and Full Left Ventricle Quantification [396]. Across these comparisons, the proposed approach maintains higher balanced accuracy and improved PR characteristics.

Table 5.5: Comparison of STARFLOW with state-of-the-art methods.

Methods	**Acc.**	**AUC**	**PR**	**Sens.**	**Spec.**	**PPV**	**NPV**
STARFLOW	87.3±6.10	0.93±0.03	0.81±0.10	82.4±12.25	92.2±1.75	72.8±11.10	95.4±3.80
Direct Quant. [410]	81.7±3.70	0.82±0.04	0.73±0.05	70.3±9.55	93.2±1.85	68.5±10.40	89.2±3.60
Direct Quant. w Att. [409]	79.5±8.20	0.82±0.07	0.73±0.12	67.6±16.20	91.4±3.30	69.4±9.30	93.5±4.00
Full Left Ven. Quant. [396]	79.1±3.75	0.78±0.07	0.69±0.08	67.0±13.25	91.2±3.50	59.5±10.80	89.0±5.70

A sensitivity analysis across heterogeneous X-ray systems indicates stable performance across acquisition devices (Table 5.6), with balanced accuracy ranging from 87.4 ± 6.0 to 90.1 ± 5.9.

Table 5.6: Sensitivity analysis: performance of STARFLOW across various X-ray systems.

Acquisition device	**Spatial resolution**	**Frame rate**	**Accuracy**	**AUC**	**Sens**	**Spec**
GE HealthCare Innova IGS 520	2.5lp/mm, 512 x 512 pixels	15	89.5 ± 6.5	0.94 ± 0.04	80.0 ± 13.9	98.9 ± 0.9
GE HealthCare Innova IGS 520	2.5lp/mm, 736 x 736 pixels	30	90.1 ± 5.9	0.95 ± 0.01	81.5 ± 7.7	98.7 ± 1.0
Philips Allura Xper FD20	3.25lp/mm, 512 x 512 pixels	60	87.5 ± 7.1	0.93 ± 0.02	77.4 ± 14.1	97.6 ± 0.5
Philips Allura Clarity	3.25lp/mm, 512 x 512 pixels	60	89.0 ± 5.4	0.91 ± 0.05	79.9 ± 11.4	98.0 ± 0.9
GE HealthCare Optima IGS 330	2.5lp/mm, 1000 x 1000 pixels	30	87.4 ± 6.0	0.95 ± 0.01	78.5 ± 14.2	96.3 ± 0.6

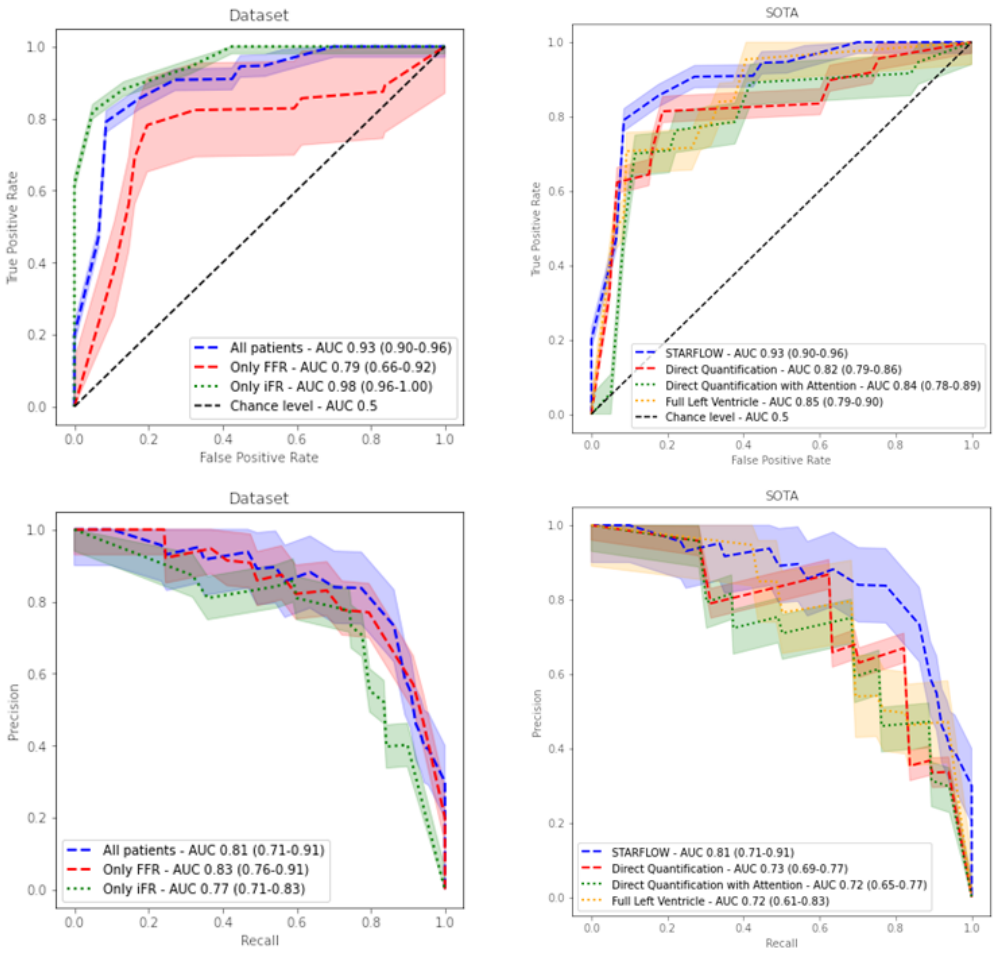


Figure 5.8: AUC-ROC (above) and precision-recall curve on a fold (below) for comparing FFR/iFR datasets (left) and comparing SOTA methods (right). Central illustration: the STARFLOW model.

5.3.8 Key-frame ablation and attention interpretability

To test whether explicit key-frame integration improves performance, we added an additional branch processing one key frame per view via a ResNet-50 and fused it late with the video representation. As shown in Table 5.7, this modification does not improve balanced accuracy, supporting the claim that the model can extract discriminative information directly from the video without manual key-frame selection.

Table 5.7: Effect of key-frame integration.

Methods	Accuracy	AUC	Sensitivity	Specificity
Ours	87.3±6.14	0.93±0.03	82.4±12.24	92.2±1.79
Ours + keyframe	87.1±7.08	0.94±0.05	82.3±10.94	91.9±1.91

We further inspected the role of the spatio-temporal attention mechanism via interpretability maps computed with M3D-CAM [51]. Fig. 5.9 illustrates that attention promotes activation on coronary segments and stenotic regions, whereas removing attention tends to increase sensitivity to non-informative structures (e.g., bones and surrounding tissues), which is consistent with less stable decision patterns.

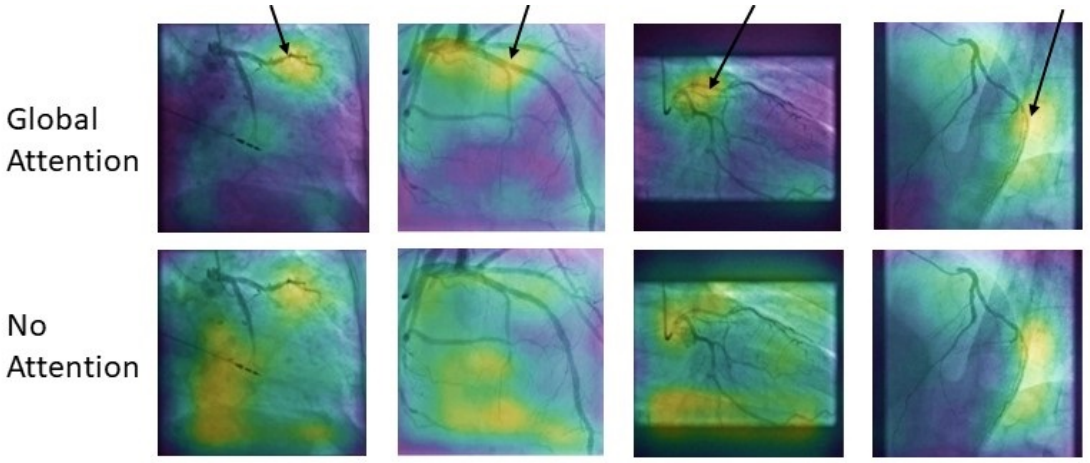


Figure 5.9: Impact of attention strategies. Interpretability maps computed through M3D-CAM [51] with attention enabled vs disabled. Yellow denotes the most activated regions, purple the least activated ones. With attention, activations concentrate on vessels and stenotic segments; without attention, activations extend to non-coronary anatomy visible in X-rays.

Overall, across multi-center acquisition settings, the proposed hybrid 3D CNN + spatio-temporal attention model achieves strong discrimination of haemodynamically significant stenoses and maintains consistent performance under heterogeneous imaging systems, while providing continuous iFR/FFR estimates that align with the invasive reference standard.

5.4 Discussion

In this work, we investigated the feasibility of deriving physiological information from routine coronary angiography to support decision-making in patients with coronary artery disease (CAD) and intermediate coronary stenoses. We developed and evaluated an AI-based system, STARFLOW, that estimates both FFR and iFR from plain coronary angiography (CA) using two angiographic projections per lesion. Methodologically, the approach couples a 3D convolutional feature extractor with spatio-temporal self-attention, enabling end-to-end learning directly from angiographic videos without manual selection of key frames or explicit modeling of vessel geometry. This design targets a practical objective: to approximate

physiological assessment while reusing data already acquired in standard workflows, potentially reducing dependence on pressure-wire instrumentation.

Across a multi-center cohort of 389 patients, STARFLOW achieved strong performance in classifying haemodynamically significant stenoses, with balanced accuracy of 87.3%, sensitivity of 82.4%, and specificity of 92.2% on the overall population assessed with invasive FFR/iFR. The model also retained satisfactory accuracy when FFR and iFR were treated as continuous values, supporting internal consistency between regression and threshold-based classification. The performance was obtained in a real-life scenario, where angiograms were acquired on different radiographic systems, with operator-chosen angulations and non-standardized views. This aspect is clinically relevant: many prior non-invasive approaches require strict acquisition constraints, complex preprocessing, or 3D reconstruction pipelines that may limit scalability in everyday practice. By contrast, STARFLOW standardizes only basic properties (spatial size, frame rate, temporal length) and learns directly from the observed contrast dynamics, facilitating translation to heterogeneous catheterization laboratory settings.

The clinical motivation for this direction remains strong. Following evidence that physiology-guided PCI improves outcomes compared with angiography-guided strategies, invasively measured FFR has become a reference standard for indicating revascularization in stable CAD [101]. However, practical barriers continue to limit routine adoption, including procedural time, costs, and the non-negligible risks and discomfort associated with intracoronary wire passage and, in the case of FFR, hyperaemia induction. The introduction of non-hyperaemic indices such as iFR mitigates some limitations, yet uptake remains modest [118]. In this context, a deployable angiography-based tool may broaden access to physiology-guided decision-making by requiring minimal additional procedure time and no dedicated probes, while preserving the option of invasive assessment when clinically indicated.

When positioned within the broader landscape of non-invasive physiology estimation, STARFLOW differs from computational and image-derived alternatives in terms of input requirements and operational complexity. Early computational models of virtual FFR from angiography demonstrated the conceptual feasibility of functional assessment from invasive imaging, but were limited by small cohorts and substantial preprocessing steps (e.g., rotational angiography export, 3D reconstruction on external workstations) [260]. At the other end of the spectrum, computational fluid dynamics (CFD) applied to coronary CT angiography has shown high diagnostic performance against invasive FFR [185, 236, 270], but is still constrained by processing times, costs, and a non-trivial learning curve, which may limit widespread near-term diffusion. Machine learning surrogates for CT-derived FFR can reduce latency, yet remain sensitive to image quality and acquisition artifacts, highlighting the dependence on consistently high-quality data for both training and deployment [61]. Compared with

these pipelines, STARFLOW leverages standard CA, does not require knowledge of vessel geometry, and was derived and evaluated under heterogeneous acquisition conditions, which supports its potential applicability in all-comers.

Several methodological observations emerge from the experiments. First, the absence of meaningful gains from explicit key-frame integration supports the hypothesis that spatio-temporal learning from the full angiographic sequence can capture the relevant contrast-flow cues without manual interaction. Second, attention-based components improved interpretability and promoted saliency over coronary structures rather than over non-informative anatomy, aligning with the objective of extracting physiology-related dynamics from complex X-ray scenes. Third, the integration of clinical covariates did not improve balanced accuracy and reduced sensitivity, despite increasing specificity. This result should be interpreted as an informative negative ablation rather than as a definitive conclusion on the clinical value of structured variables. One possible explanation is that the visual representation already encodes most of the discriminative information available in the present dataset, while the additional covariates are too coarse or weakly informative to refine borderline predictions. Another possible factor is the imbalance between the dimensionality and statistical structure of angiographic embeddings and tabular variables: without dedicated normalization, projection, or gating, late concatenation can alter the decision boundary without improving the underlying physiology-oriented representation. This behavior is especially plausible in a limited-data setting, where adding fusion parameters can increase overfitting and reduce sensitivity. Notably, one potential motivation for adding clinical features would be to account for heterogeneity in hyperaemic response and microvascular dysfunction, which can influence FFR measurements [43]. The lack of improvement suggests that either the selected covariates were insufficient to capture this variability, or that richer and more structured fusion mechanisms are required to exploit them effectively.

Future work should therefore investigate clinical-context integration through more principled multimodal designs. Possible alternatives include modality-specific encoders for tabular variables, bottleneck projections to balance feature dimensionality, gated fusion modules that learn when clinical context is reliable, FiLM-style conditioning of visual features, and cross-attention mechanisms in which clinical tokens modulate angiographic representations. A complementary option is decision-level fusion, where image-only and clinical-context models are calibrated separately and combined only at inference time. However, such strategies require larger cohorts and careful regularization, because increasing fusion complexity under limited supervision may further increase overfitting risk.

A further clinically relevant finding is the higher predictive accuracy observed among lesions assessed with iFR compared with those assessed with FFR. A plausible explanation is

that iFR is derived under resting conditions and during a diastolic wave-free period, whereas FFR depends on achieving stable maximal hyperaemia. Inter-individual variability in the response to hyperaemic agents, particularly when induced via intracoronary adenosine and in lesions supplying large myocardial territories, may increase measurement variability and reduce the predictability of the hyperaemic index from angiographic appearance alone. In our cohort, intracoronary administration was prevalent among FFR cases and LAD lesions were frequent, which may have contributed to less stable hyperaemic conditions. This hypothesis remains speculative, but it is consistent with the observed subgroup performance and highlights the importance of carefully characterizing physiology acquisition protocols when developing learning-based surrogates.

5.4.1 STARFLOW in clinical practice

From a translational standpoint, STARFLOW is envisioned as a software tool integrated into catheterization laboratory workflows. Clinicians could select or automatically retrieve two standard projections, and obtain both continuous estimates of FFR/iFR and a dichotomous classification according to established cutoffs. Deployment could be local (on a workstation equipped with a suitable GPU) or remote via a controlled cloud infrastructure, with access through a web interface. This practical framing is important because many barriers to physiology adoption are logistical rather than conceptual: operator inertia, time constraints, drug-related side effects, and probe costs are common impediments [355]. An angiography-based AI tool may lower the threshold for physiological assessment in cases where pressure-wire measurements are not performed, potentially increasing the proportion of lesions evaluated with a physiology-informed approach. At the same time, such a system should be considered as decision support, not as a replacement for invasive physiology when clinically required, and must be validated in outcome-oriented studies before influencing clinical pathways.

5.4.2 Limitations and future directions

This study has limitations that motivate further work. First, because most patients underwent invasive physiology for clinical reasons, a selection bias toward a relatively high burden of angiographic and functional disease cannot be excluded. Second, the preponderance of LAD assessments, while supporting performance in a critical vascular territory, may limit generalizability to other vessels; additional cohorts with more balanced vessel distributions are needed. Third, external validation on independent populations is currently missing. Although the use of five-fold nested cross-validation mitigates optimism bias by separating training, validation, and test splits within the available dataset, generalizability across institutions,

protocols, and patient mixes must be established prospectively. Finally, exclusions due to missing raw data or insufficiently reliable invasive traces highlight practical data-quality constraints; however, only a minority of initially screened cases lacked the minimal requirement of two views, a limitation that can be addressed operationally by incorporating this constraint into acquisition routines.

Future research directions include expanding the dataset to cover a broader spectrum of patient profiles and lesion characteristics, and conducting external and prospective validations, ideally including outcome-based endpoints to quantify the clinical risk associated with false negatives. In addition, privacy-preserving synthetic data generation may be explored both for augmentation and for enabling secure multi-center data sharing, following recent strategies in medical imaging [285]. Finally, given the modular design of STARFLOW, further work can investigate calibration across acquisition systems, improved multimodal fusion strategies for structured clinical covariates, and refined uncertainty estimation to better support clinician trust and safe deployment.

In conclusion, STARFLOW provides a non-invasive AI-based approach to derive both iFR and FFR estimates from standard coronary angiography. By avoiding hyperaemia induction, intracoronary wire passage, and complex geometric preprocessing, and by operating under heterogeneous real-world acquisition conditions, the proposed system has the potential to facilitate broader access to physiology-guided assessment of intermediate coronary stenoses and to support workflow-compatible decision-making, pending rigorous external and prospective validation.

5.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 730 hours of computation was performed on two NVIDIA Tesla T4 GPUs (thermal design power of 70 W), with total emissions estimated[2] to be 22.08 kg CO_2eq.

[2]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 6

A Convolutional-Transformer Model for FFR and iFR Assessment from Coronary Angiography

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, F Proietto Salanitri, G Bellitto, I Kavasidis, O De Filippo, M Millesimo, GM De Ferrari, M Aldinucci, D Giordano, S Palazzo, et al. A convolutional-transformer model for ffr and ifr assessment from coronary angiography. *IEEE Transactions on Medical Imaging*, 43(8):2866–2877, 2024
- Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Procedimento e sistema di elaborazione di immagini angiografiche, 2026. IT Patent 202400014434A1
- Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Angiographic image processing method and system, 2025. WO Patent 2025262634A1

6.1 Overview

Quantifying the severity of coronary stenosis from X-ray catheter angiography is the first evaluation step during invasive coronary assessment and directly influences whether revascularization should be performed [267, 183]. In routine practice, severity is often estimated visually, but this process can suffer from substantial inter-observer variability depending on operator experience, patient presentation, and the propensity to perform additional invasive measurements [120]. Clinical guidelines therefore recommend physiological assessment using FFR or iFR to determine whether an intermediate lesion is hemodynamically significant [267, 183]. In particular, stenoses with FFR below 0.80 are generally considered functionally significant [267, 170, 359, 75], while iFR uses an analogous threshold around 0.89 [267, 23, 71]; conversely, lesions above these thresholds typically do not benefit from revascularization compared to optimal medical therapy alone [267, 100]. Despite their clinical value, pressure-wire FFR/iFR measurements introduce additional time and costs and carry a low but non-negligible complication risk due to their invasive nature, motivating non-invasive alternatives that can support decision making in the cath lab.

Recent advances in deep learning have enabled strong performance across medical imaging tasks, including segmentation [293, 357], diagnosis [201, 335], and related biomedical applications [290, 83]. Alongside convolutional neural networks (CNNs), vision transformers have further advanced performance and have shown appealing properties in terms of robustness and interpretability [85, 319, 369, 414]. In coronary angiography, several methods have been proposed for automated stenosis assessment and functional inference [209, 261, 410, 409, 396]. Notably, many high-performing pipelines rely on multiple angiographic views and require selecting a *key frame* with optimal contrast and vessel delineation [409, 410]. While effective, these requirements can increase patient burden (additional acquisitions) and impose extra workflow effort, as key-frame selection is typically performed manually by cardiologists.

To address these limitations, we propose a hybrid CNN-Transformer approach for angiography based, non-invasive estimation of both FFR and iFR using **a single angiography view** and **without requiring key-frame identification**. The model is designed to infer whether a lesion is hemodynamically significant and to provide direct continuous estimates of FFR and iFR. The architecture combines the inductive bias of CNNs for local spatio-temporal representation learning, implemented via 3D convolutions, with the global modeling capabilities of a transformer encoder that employs a **spatio-temporal factorization mechanism** to redesign self-attention and capture long-range dependencies in space and time in two separate stages. This factorized formulation is motivated by the need to model complex interactions between vessel geometry, contrast-flow dynamics, and lesion morphology, while

mitigating ambiguities typical of visual inspection in intermediate lesions.

A second key component is a **multi-task, multi-branch output design** that jointly addresses stenosis assessment from two complementary perspectives: (i) classification of hemodynamic significance according to established clinical cut-offs, and (ii) direct regression of FFR and iFR values. By coupling regression and classification objectives, the model is explicitly encouraged to learn robust representations around the clinically critical FFR region near 0.80 and the iFR region near 0.89, where treatment decisions are most sensitive. This formulation also provides a practical mechanism to exploit **heterogeneously labeled datasets**, enabling supervision when targets are available either as discrete categories (threshold-based) or as continuous FFR/iFR values.

We evaluate the method on a multi-center dataset collected from multiple Italian hospitals, comprising **778 angiographic exams from 389 patients**. The proposed approach achieves state-of-the-art performance for both classification and regression, while avoiding the need for multi-view inputs or key-frame information. Beyond predictive performance, we analyze **interpretability and calibration**, showing that the model can robustly identify stenosis-relevant regions and that its prediction uncertainty correlates with the output scores, supporting more reliable clinical decision support under the intrinsic complexity of angiographic imaging data.

In summary, the main contributions are: (i) a hybrid convolutional-transformer model with factorized spatio-temporal feature extraction to capture interactions among anatomy, flow dynamics, and lesion morphology for FFR/iFR quantification; (ii) a multi-branch architecture that supports both threshold-based significance assessment and direct FFR/iFR regression, encouraging robustness around clinically relevant cut-offs and enabling training under heterogeneous labeling; (iii) an extensive multi-center experimental validation on 778 exams; and (iv) interpretability and confidence calibration analyses that support the reliability of the model's decisions compared to prior approaches.

6.2 Method

6.2.1 The proposed method

The proposed model, depicted in Fig. 6.1, consists of a cascade of spatio-temporal feature extraction blocks, based on a sequence of 3D convolution and spatio-temporal transformer layers, followed by multiple output branches, predicting a binary class on the hemodynamical significance of a stenosis (based on the thresholds suggest by the established literature [267, 359, 75, 23], i.e., 0.80 for FFR and 0.89 for iFR) and direct estimates of FFR and iFR values.

The proposed architecture is designed to make use of convolutional feature extractors, which benefit from a strong inductive bias that helps in the extraction of local visual features, and of transformer layers based on self-attention, which are instead better at finding global correlations between regions of the input data and have been shown to improve model explainability. In the following, we introduce and describe in detail each component of the model.

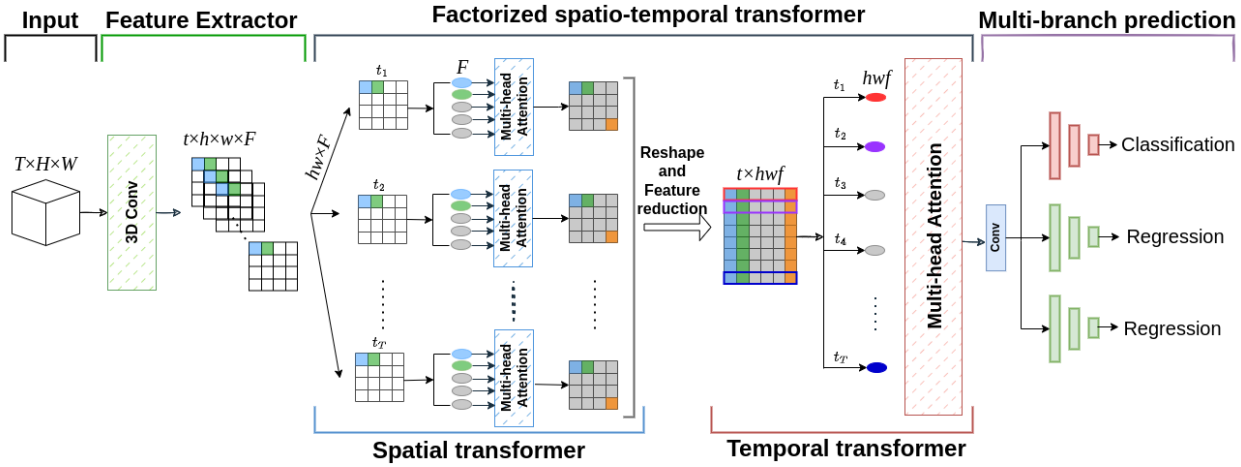


Figure 6.1: Architecture of the proposed approach for stenosis significance assessment. Input angiography videos are first processed by a pre-trained 3D convolutional model for local feature extraction. Then, self-attention layers based on transformers are employed to capture intra-relations in space and time, and the resulting intermediate representation is fed to three output branches, providing predictions as either a significance class or a regression estimate of FFR/iFR values.

6.2.2 Feature extraction through 3D convolutions

In spite of the wider and wider diffusion of transformer architectures for vision tasks, one of their main limitations stands on the lack of a strong inductive bias that can take advantage of the regular structures of real-world images. On the contrary, convolutional architectures can quickly and effectively learn distinctive and reusable patterns from visual data, but they lack a principled way to extract global features without resorting to lossy downsampling operators. For these reasons, the first processing block of our model consists of a spatio-temporal feature extractor based on 3D convolutions, able to capture meaningful patterns from video sequences, that can be later refined by more complex processes.

Given an input video sample $\mathbf{X} \in \mathbb{R}^{T \times H \times W}$, with T, H and W being, respectively, the number of frames and the height and width of each frame, we initially extract local spatio-temporal features by feeding it into a 3D convolutional neural network $\mathcal{F}$. The output of

the feature extractor is a set of F feature maps $\mathbf{H} \in \mathbb{R}^{t \times h \times w \times F}$, with t, h and w respectively smaller than T, H and W due to the downscaling effect of convolutional encoders.

In our implementation, we employ a ResNet3D [364] model as a feature extractor, pre-trained on the Kinetics-400 dataset [48] for video action recognition. This choice is in line with the motivation for employing convolutional operators only at the first stage of the model: in spite of the strong domain shift between natural videos and coronary angiographies, our usage of 3D convolutions as local feature extractors takes advantage of the generalizability of low-level features learned by pre-trained models; higher-level semantics and analysis are then carried out by the transformer blocks that follow the initial feature extraction stage. The weights of the feature extractor are fine-tuned along with the entire network at training time. In our preliminary experiments, alternative strategies such as training from scratch or freezing pre-trained weights led to low performance. Note that in order to adapt the pre-trained ResNet3D model from RGB to X-ray inputs, we simply squash the first-layer convolutional kernels by averaging over the channel dimensions.

6.2.3 Factorized spatio-temporal transformers

Local 3D features extracted by the convolutional backbone are then processed by a factorized spatio-temporal transformer, with the objective of finding relations between non-local regions within and between frames. While this operation can be performed by a purely-convolutional model through pooling layers, we argue that this approach badly fits the nature of angiographic videos for the specific task of stenosis quantification, for two main reasons. First, the spatial extent of the stenotic region often covers a small fraction of the visible area: excessive spatial downsampling inevitably results in the loss of details, which negatively affects stenosis localization (either implicit or explicit) by the model. Second, while initial temporal downsampling helps in reducing computational complexity and may take advantage of the video periodicity introduced by the cardiac cycle, temporal dimensional collapse would be needed to capture the spreading pattern of the contrast agent, whose speed may provide essential clues on both blood flow and vessel geometry; however, collapsing the temporal dimension to capture global patterns makes it impossible to extract time-varying dynamics.

To overcome these limitations, we employ a spatio-temporal transformer to process local 3D features and extract higher-level task-specific features, while retaining sufficient spatial and temporal resolution. Architecturally, we factorize the transformer into separate spatial and temporal modules, for more efficient computation [14].

Both modules internally apply scaled dot-product multi-head self-attention [373]. Let $\mathbf{T} \in \mathbb{R}^{N \times F}$ be a set of feature vectors, packed as matrix rows (F being feature dimensionality).

A transformer layer produces a new set of feature $\mathbf{U} \in \mathbb{R}^{N \times F}$ as:

$$\mathbf{T}' = \mathrm{LN}\left(\mathrm{MA}\left(\mathbf{T}\right) + \mathbf{T}\right) \tag{6.1}$$

$$\mathbf{U} = \mathrm{LN}\left(\mathrm{MLP}\left(\mathbf{T}'\right) + \mathbf{T}'\right) \tag{6.2}$$

where LN is layer normalization [18], MLP is a two-layer multi-layer perception, with input and output sizes equal to F and a hidden layer with d_{MLP} neurons with ReLU activation, and MA is the multi-head self-attention operator, defined in the following. Note that the input and output dimensions of the transformer layer are unchanged (equal to $N \times F$), allowing for arbitrary sequences of layers. In each layer, the self-attention function A computes query, key, and value vectors for each feature vector in $\mathbf{T}$ by linear projection into $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{N \times d}$, with d as the feature dimension. The projection matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V \in \mathbb{R}^{F \times d}$ are learnable. The output for each $\mathbf{T}$ row is a linear combination of $\mathbf{V}$ rows, weighted by dot-product similarity calculated as $\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)$. The multi-head self-attention function MA receives h sets of projection matrices $\{\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V\}_{i=1}^h$. Outputs from the self-attention function are concatenated and linearly projected using $\mathbf{W}^O \in \mathbb{R}^{F \times F}$ to produce the final output. The attention feature dimensionality d is simplified to F/h.

In our model, the spatial transformer module receives convolutional features $\mathbf{H} \in \mathbb{R}^{t \times h \times w \times F}$, and processes each frame $\mathbf{H}_i \in \mathbb{R}^{h \times w \times F}$ independently. The height and width dimensions are grouped, in order to reshape each $\mathbf{H}_i$ as $\mathbf{T}_i \in \mathbb{R}^{N_{\mathrm{space}} \times F}$ (with $N_{\mathrm{space}} = wh$), with dimensions suitable for multi-head self-attention. The outputs of all spatial transformers are then reshaped and concatenated back as new features $\mathbf{H}_{\mathrm{space}} \in \mathbb{R}^{t \times h \times w \times F}$.

The temporal transformer module applies a similar operation on $\mathbf{H}_{\mathrm{space}}$, with the difference that attention is computed between concatenated frame-level feature vectors. First, in order to reduce feature dimensionality (which would explode due to concatenation), we project each spatial feature vector from F to a lower size f; then, all features within each frame are concatenated, resulting in a tensor $\mathbf{H}'_{\mathrm{space}} \in \mathbb{R}^{N_{\mathrm{time}} \times hwf}$ (with $N_{\mathrm{time}} = t$), which can be fed to the temporal transformer layers, producing the output tensor $\mathbf{H}_{\mathrm{time}} \in \mathbb{R}^{t \times h \times w \times f}$ (after reshaping).

In our implementation, both spatial and temporal transformers include two layers each, with $h = 4$ attention heads per layer.

6.2.4 Multi-branch prediction

The output $\mathbf{H}_{\mathrm{time}}$ of the temporal transformer undergoes a dimensionality reduction step by means of two layers of 3D convolutions with spatio-temporal stride equal to 2, thus reducing each dimension by a factor of four: the resulting tensor is then flattened into a

vector $\mathbf{V} \in \mathbb{R}^{fthw/64}$.

The multi-branch prediction module receives $\mathbf{V}$ and feeds it into three separate two-layer MLPs, with batch normalization and ReLU activations. The hidden layer of each MLP halves the number of input neurons (from 512 to 256) and the output layer projects to a single scalar (from 256 to 1). Thus, the final output of the layer is a set of scalars $\hat{y}_c$ (constrained between 0 and 1 through a sigmoid activation), $\hat{y}_{\text{FFR}}$ and $\hat{y}_{\text{iFR}}$, respectively predicting a binary class of stenosis significance and the direct estimates of FFR and iFR values.

A practical issue in this task is that FFR and iFR values are not simultaneously available for all samples. This reflects routine clinical practice: once one physiological index is acquired and considered sufficient for decision-making, the other one is often not measured. Therefore, the classification target can be derived whenever at least one physiological measurement is available, using the standard clinical thresholds, whereas the regression targets are only available for the corresponding measured index. To avoid propagating gradients from missing regression targets, we introduce binary availability masks m_i^{FFR} and m_i^{iFR} for each sample i, where the mask is equal to 1 if the corresponding target is available and 0 otherwise.

Given ground-truth values y_c, y_{FFR} and y_{iFR}, we define the training objective $\mathcal{L}$ as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{FFR}} + \mathcal{L}_{\text{iFR}} \\ &= -y_c \log \hat{y}_c - (1 - y_c) \log (1 - \hat{y}_c) \\ &\quad + \lambda m^{\text{FFR}} \left| y_{\text{FFR}} - \hat{y}_{\text{FFR}} \right| + \lambda m^{\text{iFR}} \left| y_{\text{iFR}} - \hat{y}_{\text{iFR}} \right|, \end{aligned} \tag{6.3}$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy classification loss, $\mathcal{L}_{\text{FFR}}$ and $\mathcal{L}_{\text{iFR}}$ are masked L_1 regression losses, and λ is a weighing factor between classification and regression terms. In this formulation, unavailable FFR/iFR values do not contribute to the corresponding regression branch. The classification branch, instead, remains supervised on the largest possible cohort because both FFR and iFR can be mapped to the same clinical decision of haemodynamic significance using established thresholds.

6.3 Experimental Results

6.3.1 Dataset

We use a private dataset comprising 778 coronary angiographies, retrospectively collected between January 2020 and January 2022. The dataset encompasses two examinations (two examples are reported in Fig. 6.2) from 389 patients diagnosed with CCS and ACS, distributed across five hospitals as per Tab. 6.2 (first two columns). Patients underwent

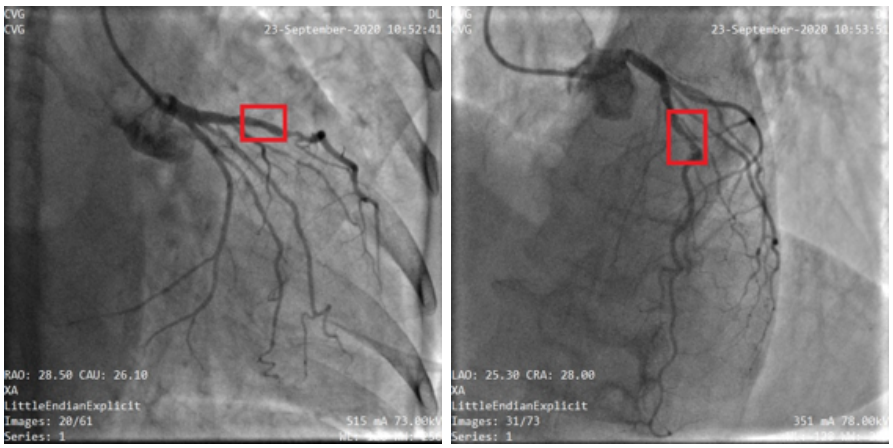


Figure 6.2: Examples of the two views per subject. Bounding boxes indicate major stenoses as identified by two expert radiologists.

invasive physiological assessment of intermediate single coronary stenoses (diameter of stenosis $\geq 30\%$ and $\leq 90\%$ at angiographic visual estimation) through the measurement of FFR, iFR or both. Among the 389 patients, 303 were men and 86 women, and the mean age was 67.9 $\pm$ 9.61. IRB protocol number is 0092163.

Each angiography was evaluated by two expert cardiologists (for each hospital) through invasive physiological assessment of intermediate coronary stenosis with iFR, FFR, or both. In particular, FFR values are provided for 251 patients (64.5%), iFR values for 228 patients (58.6%); a set of 90 patients (23.1%) include both. For each exam, the major stenosis was identified by radiologists and labelled as hemodynamically significant if FFR is lower than 0.80 [267, 170, 359, 75] or if iFR is lower than 0.89 [267, 23, 71]. As a result, 94 patients (24.4%) are labeled as positive, while the remaining cases are negative. Coronary angiography and physiological measurements were performed following standardized clinical practice. Key frames were annotated by two expert cardiologists.

Since the angiographies were collected with different machines and practices available in the 5 involved hospitals, their spatial sizes and frame rates vary, respectively, from 512×512 to 1024×1024 and from 15 fps to 60 fps. Given these differences, all samples were resized to 256x256 pixel ($H \times W$) and 15 fps (downsampling when needed). All collected videos were cut to a length of 60 frames by selecting the 4 central seconds of the videos. This choice is motivated by the typical range of angiographies (about 3.5 s), in order to be strategically centered around the region of interest with the highest level of perfusion. In cases the angiography video was shorter than 4 seconds, we employ padding by replicating the first and the last frames up to 4 seconds.
In all the single-view experiments, we use only one random view per patient (either in training or validation/test) excluding the other remaining views.

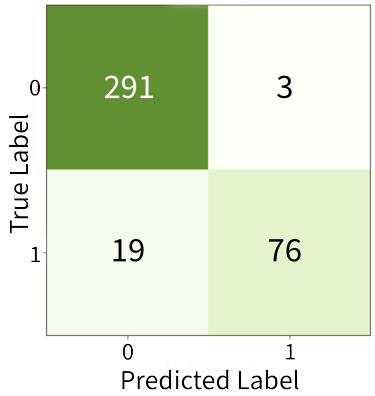


Figure 6.3: Cumulative confusion matrix over the five tested folds.

6.3.2 Training and Evaluation Procedure

Table 6.1: Classification performance comparison of our model using different input data modalities.

	Modality	Accuracy(%)	AUC	Sensitivity(%)	Specificity(%)
Single-view	3D	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	80.0 ± 13.9	$\mathbf{98.9 \pm 0.9}$
Single-view + key frame	2D+3D	86.4 ± 8.8	0.86 ± 0.09	79.4 ± 14.6	93.4 ± 7.4
Multi-view	3D	89.3 ± 7.0	0.95 ± 0.04	$\mathbf{80.9 \pm 13.9}$	97.2 ± 2.0
Multi-view + key frames	2D+3D	89.1 ± 6.1	0.94 ± 0.05	80.0 ± 11.6	96.2 ± 2.1

We train our model to minimize the multi-task loss $\mathcal{L}$ from (6.3), through gradient descent with the AdamW optimizer [188] (learning rate: 1e-5, batch size: 8), for a total of 300 epochs. Input X-ray angiography videos were standardized and augmented through: 1) 2D horizontal and vertical flipping and 90° rotations (identically applied to all frames of the same video), and 2) temporal sampling rate augmentation by reducing arbitrarily the frame rate for a more effective learning of flow velocity. Furthermore, to deal with the overrepresentation of negative cases, we employed random oversampling by randomly duplicating instances from the positive class.

At training time, the two regression branches are not necessarily both activated, since not all samples are provided with both FFR and iFR values: as a result, L_1 loss terms $\mathcal{L}_{\text{FFR}}$ and $\mathcal{L}_{\text{iFR}}$ are occasionally ignored when the corresponding target is not present. To cope with class imbalance, as classification loss, we employed balanced binary cross entropy$\mathcal{L}_{\text{BCE}}$ defined as follows:

$$BCE(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[\alpha \cdot y_i \cdot \log(p_i) + (1-\alpha) \cdot (1-y_i) \cdot \log(1-p_i)] \tag{6.4}$$

where y is the true target value (actual label), p_i is the predicted probability for the positive class. N is the number of samples in the dataset and α the weight or factor that balances the contribution of positive and negative classes. In our case, it is set to the ratio of the number of positive samples (less represented class) to the total number of samples, ensuring that both classes have equal influence on the loss.

BCE is optimized as follows: the target class, representing hemodynamical stenosis significance, is based on the availability of FFR and iFR values and on the corresponding thresholds. If both are available, FFR is used to set the classification target. The λ weighing factor in (6.3) is empirically set to 10, based on the relative magnitudes of cross-entropy and L_1 losses.

We employ a ResNet3D architecture for feature extraction. Given a $T \times H \times W = 60 \times 256 \times 256$ input, the output feature volume is $t \times h \times w \times F = 8 \times 16 \times 16 \times 128$. Feature size F is obtained by reducing ResNet3D's output features from 512 to 128 through a linear projection. Output features from the spatial transformer are further projected from $F = 128$ to $f = 16$ to prevent feature explosion due to temporal concatenation. The size of the input to the multi-branch module is therefore $ftwh/64 = 512$. Finally, the multi-branch prediction module comprises three identical blocks, each one composed by a first fully connected layer with 256 neurons followed by ReLU and dropout ($p = 0.1$) and a final classification layer with two neurons.

We carry out model evaluation for the proposed approach and for state-of-the-art methods on the classification task, reporting balanced accuracy (to account for class imbalance), AUC, sensitivity and specificity. For our approach, we also report regression accuracy on FFR and iFR scores as MSE and MAE. Performance metrics of the proposed approach and the methods under comparison are computed through 5-fold stratified nested cross-validation. For each experiment fold, 60% of the dataset is used for training, 20% for validation and 20% for test. Validation accuracy is used to select the training epoch at which test performance is evaluated: performance metrics are then averaged over all test folds. Furthermore, the split of the dataset into training, validation, and test sets was performed on a per-patient basis, in order to guarantee that angiograms from a single patient are never unintentionally split between the training and test sets, thus reducing any potential overestimation of generation performance due to correlation among samples.

Experiments are carried out on two NVIDIA Tesla T4 GPUs using automatic mixed precision for training. The proposed approach was implemented in PyTorch and MONAI; all code is available at https://github.com/perceivelab/conv-transf-ffr-ifr-assessment.

Table 6.2: Per-hospital classification performance obtained by the proposed approach. Performance is consistent across the five centers despite the high data imbalance.

Hospital	**N_Patients**	**TP**	**FN**	**FP**	**TN**	**Acc.(%)**	**Sens.(%)**	**Spec.(%)**
1	90	10	3	0	77	88.46	76.92	100.00
2	121	8	2	1	110	89.55	80.00	99.10
3	58	33	8	1	16	87.30	80.50	94.11
4	80	14	4	1	61	88.08	77.78	98.39
5	40	11	2	0	27	92.31	84.61	100.00

Table 6.3: Classification performance comparison with the state of the art. For each model, we report the input modality as well as balanced accuracy (*Acc.*), AUC, sensitivity (*Sens.*) and specificity (*Spec.*), computed through 5-fold nested cross-validation. S stands for single view, i.e., only one view per patient is used, while M stands for multiview, i.e., at least two views per patient are employed.

	Methods	**View**	**Input**	**Acc. (%)**	**AUC**	**Sens. (%)**	**Spec. (%)**
CNN	**S3D** [395]	S	2D+t	79.5 ± 5.5	0.93 ± 0.03	65.4 ± 14.2	93.6 ± 4.8
CNN	**ResNet3D** [364]	S	2D+t	77.9 ± 5.1	0.85 ± 0.04	64.0 ± 9.1	91.8 ± 1.8
CNN	**MVCNN** [350]	M	2D	84.5 ± 6.1	0.85 ± 0.07	76.8 ± 10.2	92.2 ± 2.2
CNN	**GVCNN** [103]	M	2D	78.4 ± 4.5	0.87 ± 0.05	71.0 ± 7.3	85.8 ± 4.4
Transformer	**ViT-B_16** [85]	S	2D	76.8 ± 6.1	0.74 ± 0.11	69.0 ± 14.1	84.6 ± 5.9
Transformer	**ViT-3D** [319]	S	2D+t	73.7 ± 3.7	0.78 ± 0.06	63.0 ± 7.4	84.4 ± 1.2
Transformer	**Swin Transformer** [214]	S	2D	63.8 ± 9.7	0.70 ± 0.11	41.0 ± 27.2	86.6 ± 8.2
Transformer	**Video Swin** [216]	S	2D+t	81.2 ± 7.0	0.92 ± 0.06	65.2 ± 14.2	95.6 ± 0.5
Hybrid	**DMQCA** [410]	M	2D-(2D+t)	81.7 ± 4.2	0.82 ± 0.04	70.2 ± 9.1	93.2 ± 2.1
Hybrid	**DMTRL** [396]	S	(2D+t)	79.1 ± 4.2	0.85 ± 0.06	67.0 ± 7.5	91.2 ± 3.9
Hybrid	**HEAL** [409]	M	2D-(2D+t)	79.5 ± 5.9	0.84 ± 0.06	67.6 ± 10.6	91.4 ± 3.7
	Proposed method	S	2D+t	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

6.3.3 Effect of input data modality

Most of existing methods for stenosis quantification, such as DMQCA [410], HEAL [409], DMTRL [396], employ multiple views and key frame information. Thus, our first battery of experiments aims at evaluating how input data modality affects model performance. In particular, we compare our method with variants thereof, using: 1) an additional convolutional branch (consisting of a ResNet18 network) that extracts 2D features from the key frame and concatenates them, before multi-task prediction, with the features provided by the spatio-temporal transformer; 2) two views (as existing methods, i.e., DMQCA [410], HEAL [409]): in this case, the model backbone (CNN and transformer) is shared across the views and feature concatenation is carried out before multi-task prediction; 3) two views in addition to two key frames (one for each view), with each keyframe processed as presented in item 1. Table 6.1 reports the results of this comparison: our model outperforms multi-view variants

on three out of four metrics, while being very close on the fourth. It is interesting to notice that adding key frame information seems to hinder performance. This may be due to the fact that, as also shown by other experiments presented in the following, the model mostly focuses on temporal information for its predictions; spatial-only features from a single frame may thus introduce noise by increasing dimensionality with no informative contribution. Multi-view inputs, instead, do not seem to significantly harm performance, but the lack of improvements demonstrates our method's capability to perform well with a single view, avoiding extra burdens on physicians and patients.

We further delve into the performance of the single-view model by inspecting the cumulative confusion matrix among the five folds, shown in Fig. 6.3 and demonstrating very good specificity and satisfactory sensitivity. In order to investigate possible bias in the data and, consequently, in our approach, we also compute individual performance for each hospital. In particular, we use data from five different hospitals with different number of patients per hospital and different acquisition modality and the obtained results are reported in Table 6.2. It can be noted that, despite the unbalanced data, performance is consistent across hospitals, thus suggesting the no bias affects the yielded performance.
Furthermore, sensitivity is worse than sensitivity and this is possible due to the hard-defined clinical cut-off (0.8) for defining the positive and negative cases. This limitation is, however, mitigated by the low error (reported later) in terms of FFR regression that allows for personalized intervention.

6.3.4 Comparison with the state of the art

We then compare our performance to that achieved by state-of-the-art models. Specifically, this evaluation includes the following architectures: *CNN-based models* (S3D [395], ResNet_3D-18 [364], GVCNN [103] and MVCNN [350]); *vision transformer models* (ViT-B_16 [85], ViT-3D [319] and Video Swin Transformer [216]); hybrid models, specifically designed for angiography video analysis, that combine features from both CNN and transformer architectures (DMTRL [396], DMQCA [410] and HEAL [409][1]).

Results, shown in Table 6.3, report significantly higher performance by our model compared to the state of the art, on all metrics. Remarkably, our approach shows a gain of over ten percent points on the sensitivity score, which is the most significant metric for clinical validation of automated angiography analysis tools [307]. Moreover, this analysis confirms that our method yields better performance with fewer input data, as it only requires a single-view video, while, for instance, DMQCA [410] and HEAL [409] employ two views

[1]The authors of these approaches did not release source code, thus the results refer to our implementation thereof. For these methods, model selection was carried out through grid-search with cross-validation.

(requiring physicians and patients to carry out two exams on patients) and the key frame (to be manually identified by physicians). Fig. 6.4 reports the comparison in terms of ROC and precision-recall curves between our approach and the state-of-the-art methods specifically designed for stenosis quantification, i.e., DMQCA [410], HEAL [409] and DMTRL [396].

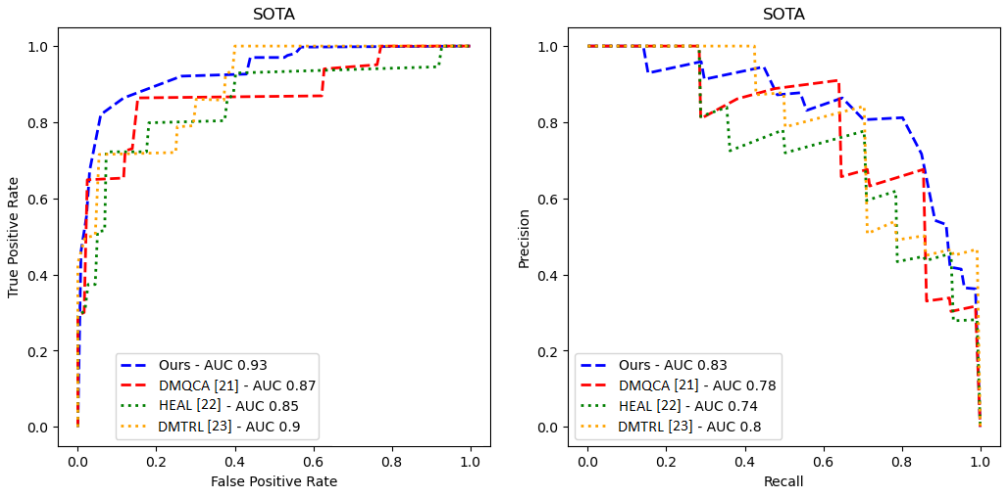


Figure 6.4: ROC (left) and Precision-Recall (right) curves comparison between our approach and state-of-the-art methods.

We evaluate our approach also in terms of interpretability and calibration to assess the reliability of the model. As for the calibration, we compute the Brier Score and compare it with the ones obtained by other models performing stenosis quantification. Results in Table 6.4 shows that our approach is more reliable for clinical applications than existing ones, as it is able to associate each prediction with a more accurate confidence score, reducing the probability of major mistakes (especially in case of false positives) and providing a more interpretable decision to physicians. To further enhance trust in our model, we also evaluate interpretability maps to assess whether our model focuses on the major stenosis, as identified by radiologists, to make its predictions. For this analysis, we employ M3D-CAM [119, 51] to compute interpretability maps for our model's and state-of-the-art models' decisions. Results, shown in Fig. 6.5, indicate that the proposed approach generally attends to the entire vessel structures, involved in the flow contrast over time, placing a particular emphasis on the major stenosis (as identified by cardiologists). In contrast, this level of comprehensive attention to vessel structures does not consistently manifest in the three methods under comparison. This observation underscores the capacity of the proposed approach to learn spatio-temporal features associated with flow dynamics, vessel geometry, and lesion morphology.

Table 6.4: Calibration performance in terms of Brier score; lower values indicate better calibration.

Methods	**Brier Score** (↓)
DMQCA [410]	0.145 ± 0.032
HEAL [409]	0.125 ± 0.041
DMTRL [396]	0.154 ± 0.058
Proposed method	0.089 ± 0.047

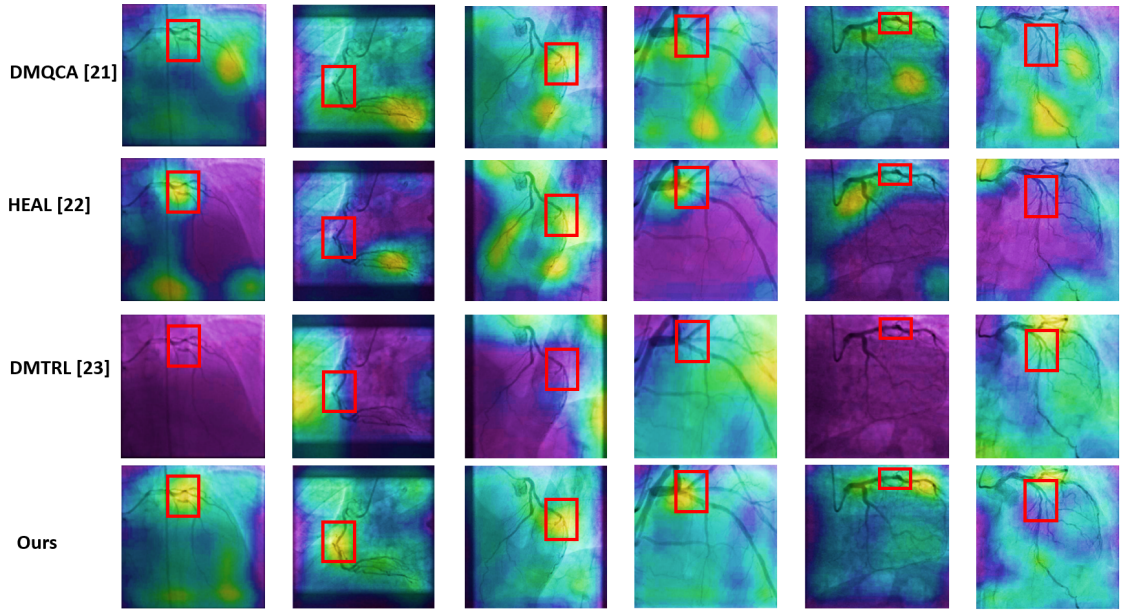


Figure 6.5: Interpretability maps, computed by M3D-CAM [119, 51], of our method (last row) and DMQCA [410], HEAL [409] and DMTRL [396] (first, second, and third row, respectively). For each image, we also report, as a red bounding box, the location of the major stenosis, as identified by cardiologists. In each map, the yellow parts are the most activated ones, while the purple parts are the least activated ones.

6.3.5 Ablation Studies

We then carry out ablation studies to validate our architectural design choices.

Starting from a baseline model, consisting of a 3D convolutional network (ResNet3D) with a single classification branch, we add, first individually then together, the spatio-temporal transformer and the multi-task branch (which introduces FFR/iFR regression). Results, given in Table 6.5, indicate how each block contributes to enhance performance. It is interesting to note that the major gain is obtained when including the multi-branch module with multi-task loss: in particular, sensitivity increases by more than five percent points with respect to the hybrid convolutional-transformer network alone.

Table 6.5: Ablation on architecture modules. Results refer to classification performance.

	Acc.(%)	**AUC**	**Sens.(%)**	**Spec.(%)**
3D CNN	77.9 ± 5.1	0.85 ± 0.04	64.0 ± 9.1	91.8 ± 1.8
+ multi-branch prediction	87.9 ± 6.5	0.94 ± 0.05	78.5 ± 14.7	96.4 ± 1.2
+ factorized spatio-temporal transf.	86.6 ± 8.1	0.94 ± 0.07	74.9 ± 15.4	96.0 ± 1.5
+ multi-branch prediction	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

Table 6.6: Ablation on FFR-iFR regression on a subset of the overall dataset. Results refer to classification performance.

	Acc.(%)	**AUC**	**Sens.(%)**	**Spec.(%)**
Baseline - 3D CNN	76.6	0.71	50.0	97.2
+ FFR pred	93.3	0.93	90.9	95.7
+ iFR pred	89.8	0.92	81.8	97.8
+ FFR and iFR pred	89.9	0.92	81.8	97.9

Thus, FFR/iFR regression emerges as the optimal approach, delivering superior accuracy. Subsequently, we evaluate whether FFR or iFR yields the highest gain in classification accuracy. However, the available dataset lacks both FFR and iFR values for all exams. Consequently, we perform this evaluation on a subset of 81 cases from the original dataset, where data includes both values. FFR and iFR regression employ factorized spatio-temporal transformers and multi-branch prediction, as previously presented. Due to the smaller size of this subset, we compute results based on 60/20/20 random splits for training/validation/test, and report test values at the lowest validation loss.

The results, presented in Table 6.6, affirm the following key findings: 1) Employing regression on either FFR or iFR results in a noteworthy enhancement (exceeding 15 percentage points) compared to the baseline, even with a more limited dataset; 2) FFR emerges as more reliable than iFR; specifically, when iFR is utilized, performance slightly lags behind that of FFR but still surpasses the baseline by a considerable margin. Despite FFR's higher reliability, we opted to include iFR in the analysis for the entire dataset. This decision stems from the unavailability of FFR estimates for all exams and given the observed superior accuracy of iFR regression compared to the baselines.

The proposed model uses 60 frames to provide the network with enough data to learn flow dynamics. This choice aims to empower the model in capturing and effectively discerning subtle spatio-temporal patterns within the data. To underscore the significance of this design choice, we compare our model's performance under three different input additional video sequence lengths, namely, 15, 30, and 45 frames. As reported in Table 6.7, the impact of the frames number on performance is clear. With only 15 frames, the model exhibits significantly lower performance compared to the 60-frame configuration. We also observe a performance

gain as we increase the number of frames, confirming that increasing the temporal context helps the model to learn the motion dynamics associated to vessel geometry and lesion morphology.

Table 6.7: Ablation on number of frames. Results refer to classification performance.

#	**Accuracy(%)**	**AUC**	**Sensitivity(%)**	**Specificity(%)**
15	80.8 ± 6.4	0.90 ± 0.09	71.3 ± 11.6	90.2 ± 3.1
30	81.5 ± 5.5	0.93 ± 0.05	71.0 ± 10.2	91.9 ± 6.1
45	84.6 ± 11.6	0.95 ± 0.04	73.2 ± 13.0	96.0 ± 1.2
60	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

The positive impact of multi-branch prediction to our model suggests that regressing continuous values of FFR and iFR tends to better regularize our model, compared to using classification alone (though based on the same FFR/iFR values). To quantify how well our model predicts FFR and iFR, we compute MAE and MSE metrics, reported in Table 6.8. In order to better investigate the behavior of the model, Fig. 6.6 plots MAE with standard deviation values computed on different intervals of the FFR and iFR ranges, showing that the model is more precise around thresholds for hemodynamical significance for both iFR and FFR. This directly reflects on improved classification of hemodynamical stenosis significance. Furthermore, we also conduct an analysis of our model's performance through Bland-Altman and scatter plots, shown in Fig. 6.7. The Bland-Altman analysis indicates a strong agreement between the values predicted by our model and those measured with a mean difference of zero, indicating that there is no systematic bias. Scatter plots, instead, demonstrate that the lowest errors and the smaller standard deviations are consistently observed around clinically-relevant thresholds (0.8 for FFR, 0.89 for iFR). These findings collectively substantiate our claim on the role of multitask learning in forcing the model to improve the assessment around the clinical cut-off for FFR, as it is critical for personalized intervention.

Table 6.8: FFR and iFR regression performance of the proposed model.

	MSE	**MAE**
FFR regression	0.036 ± 0.007	0.063 ± 0.008
iFR regression	0.025 ± 0.006	0.045 ± 0.008

Finally, we investigate the impact of the attention mechanism employed within the spatio-temporal transformer encoder. In particular, we evaluate the performance of our model when using a) *global attention*, i.e., each location in the feature volume attends to all other locations in space and time; b) *temporal attention*, i.e., each location attends the same spatial location

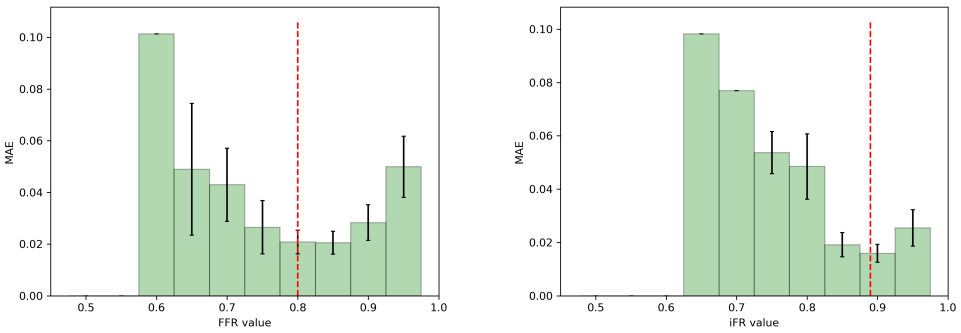
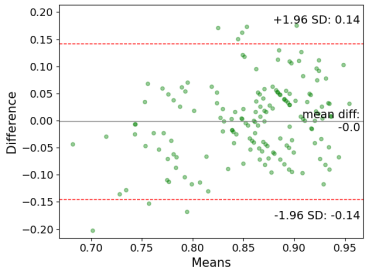
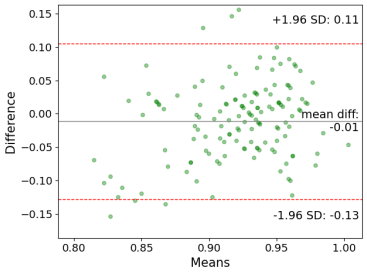


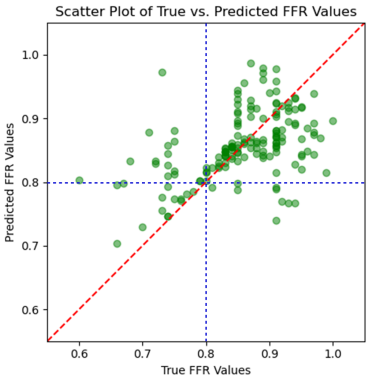
Figure 6.6: MAE at different FFR and iFR range intervals. Blue bars measure the regression MAE computed at each interval of the FFR/iFR ranges. Standard deviation on each interval is reported as an error bar. Vertical red lines correspond to thresholds for FFR and iFR, below which a stenosis is considered to be hemodynamically significant.



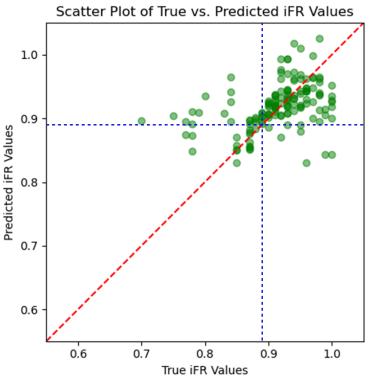
(a) Bland-Altman plot for FFR.



(b) Bland-Altman plot for iFR.



(c) Scatter plot for FFR.



(d) Scatter plot for iFR.

Figure 6.7: (First row) Bland-Altman plot between the predicted and the reference values for FFR and iFR. There is agreement between the two sets of values with 0 mean, suggesting no systematic bias in the evaluation. (Second row) Scatter plots for FFR and iFR demonstrate that the lowest errors and the smaller standard deviations are consistently observed around clinically-relevant thresholds (0.8 for FFR, 0.89 for iFR, depicted in blue dashed lines).

along the time dimension; c) *factorized spatio-temporal attention* attention is first computed on each frame independently, overall spatial locations, and then along the temporal dimension over frame-level features. The different attention strategies are illustrated in Fig. 6.8. We also compare the multi-head attention in transformers with other attention mechanisms, such as the one based on non-local blocks [383]. Also in this case, we developed the three strategies (global, temporal and factorized spatio-temporal) mentioned earlier.

Table 6.9 shows our transformer-based method yields better results than the ones obtained with non-local blocks. Among the transformer-based attention strategies, temporal attention appears to be the main responsible for improving performance; spatial information, while marginally improving accuracy when combined to temporal features in the factorized approach, is actually detriment in the global attention setting, which may be due to a difficulty by the model in learning correlation patterns over the whole spatio-temporal volume. This result, demonstrating the critical importance of temporal information in coronary angiographies, also supports our original motivation for including transformer layers for high-level analysis, in order to avoid temporal downsampling which would lead to the loss of essential information. The quantitative improvement in performance also reflects on the interpretability maps, as illustrated in Fig. 6.9.

Table 6.9: Ablation on attention strategies.

	Acc.(%)	**AUC**	**Sens.(%)**	**Spec.(%)**
Non-local block-based attention				
Global	61.9 ± 5.60	0.84 ± 0.11	25.2 ± 11.7	98.6 ± 0.8
Temporal	79.3 ± 1.18	0.89 ± 0.06	64.3 ± 19.6	94.3 ± 5.4
Factorized spatio-temporal	84.0 ± 8.40	0.91 ± 0.05	73.2 ± 17.8	94.8 ± 5.2
Transformer-based attention				
Global	83.5 ± 8.1	0.93 ± 0.03	74.6 ± 16.4	92.3 ± 1.6
Temporal	87.8 ± 7.4	0.95 ± 0.02	80.7 ± 13.8	94.8 ± 1.1
Factorized spatio-temporal	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

6.4 Discussion

Diagnosing coronary artery conditions, especially in terms of predicting functional metrics like FFR or iFR, presents significant challenges for clinicians. In this work, we present a novel hybrid convolutional-transformer method for coronary stenosis quantification through the estimation of FFR and iFR values. Unlike other methods from the established literature, our approach requires neither multiple angiography views as input, nor the identification of a key frame, thus reducing the burden on both medical staff and patients. Instead, we show

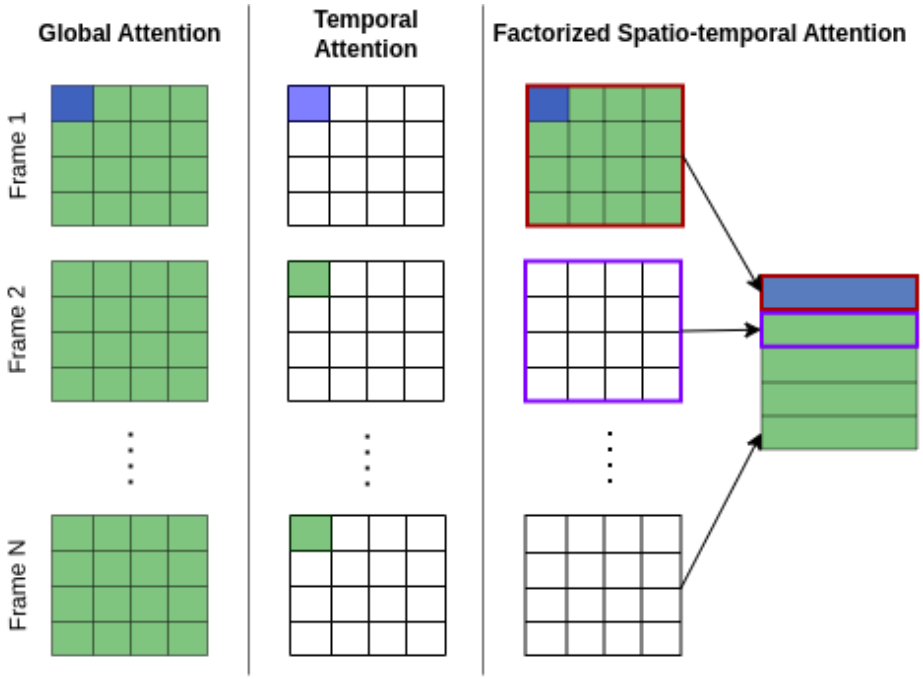


Figure 6.8: Attention mechanisms. *Global attention*: each spatio-temporal location (in blue) attends to all other locations (in green). *Temporal attention*: each location (in blue) attends the same spatial location along the time dimension (in green). *Factorized spatio-temporal attention*: attention is first computed on each frame independently, and then along the temporal dimension over frame-level features.

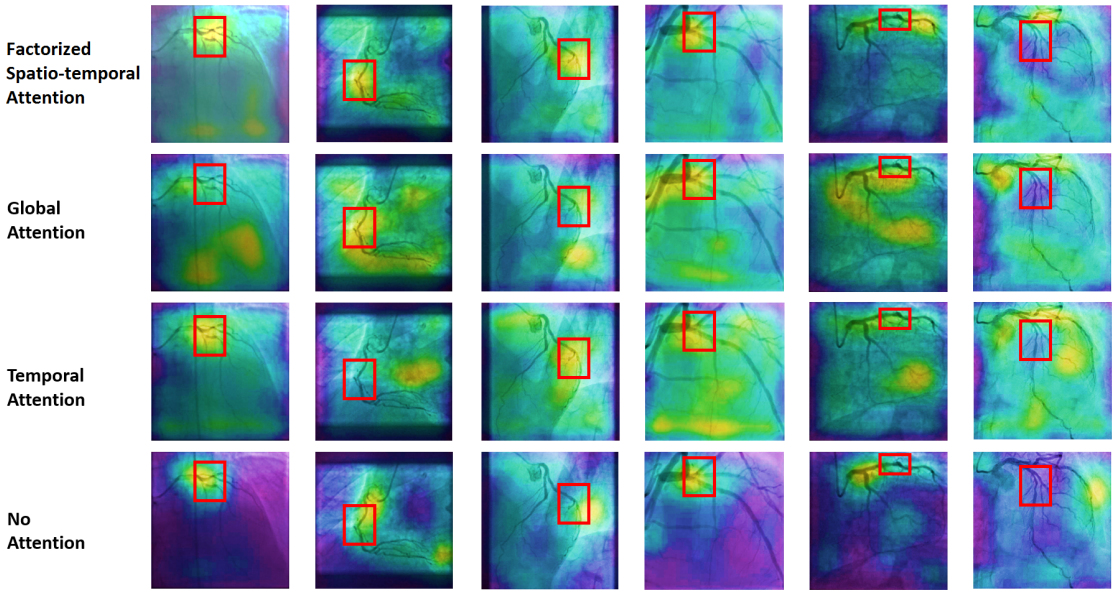


Figure 6.9: Interpretability maps, computed through M3D-CAM [119, 51], when using different attention strategies. For each image, we also report, as a red bounding box, the major stenosis as identified by clinicians. In each part, yellow regions correspond to the most activated areas, whereas purple regions correspond to the least activated ones.

that the combination of convolutional and transformer-based architectures, together with a multi-branch prediction approach for estimating stenosis hemodynamical significance (posed as a binary classification problem) as well as direct FFR/iFR scores, allows our method to outperform state-of-the-art approaches that require multiple views and/or key frame information.

The method was tested on a dataset with 389 cases, which represents the largest dataset employed so far in works for automated FFR assessment, reaching an accuracy of about 90%, a sensitivity of 80% and a specificity over 95%. These standard performance metrics, coupled with the Brier score for probabilistic predictions and model interpretability maps, affirm the model's reliability and its alignment with clinically relevant features. The extensive evaluation enhances trust and demonstrates, despite the limited dataset size, the effectiveness of the devised countermeasures to discourage overfitting.

We argue that the main reason for the success of our approach stands in its independence from a key frame. In contrast to relying on key frames that capture specific moments in the cardiac cycle, our method learns blood flow dynamics throughout the entire cycle. This forces the model to capture subtle flow pattern changes that may be missed by existing morphology-centric methods or may not be apparent to the naked eye of clinicians. Indeed, coronary artery conditions require a holistic understanding of complex interactions between factors such as vessel geometry, blood flow dynamics, and lesion morphology. Thanks to its spatio-temporal focus, the proposed method appears effective at modeling these interactions, while existing methods and visual assessment by physicians may struggle to capture them consistently. Moreover, relying primarily on the visual assessment of stenosis morphology, as human clinicians do, is a main source of inter-observer variability [121] and reduced reproducibility, while providing quantitative measures on flow-related parameters can improve consistency in the assessment.

An important aspect of the study is the partial availability of FFR and iFR labels across the cohort. This limitation affects the regression component because each regression branch is optimized only on the subset of samples for which the corresponding continuous index is available. As a consequence, the direct estimation of FFR and iFR may be less statistically stable than the binary classification task, and regression performance should be interpreted with this constraint in mind. However, this does not invalidate the classification analysis. From a clinical standpoint, both FFR and iFR can be mapped to a common decision of haemodynamic significance using established thresholds, which allows the classification branch to exploit a larger effective cohort than either regression branch alone. The masked regression formulation further ensures that missing continuous targets do not introduce incorrect gradients or penalize the model. In this sense, the multi-branch design reflects a real

clinical condition: FFR and iFR are not always acquired together because one measurement can already be sufficient for decision-making. The proposed framework therefore trades complete paired regression supervision for broader classification supervision, which is coherent with threshold-driven revascularization decisions.

In future research endeavors, we plan to carry out a more extensive assessment including more data from different geographical areas to enhance the applicability of the proposed approach. Future datasets with systematically paired FFR/iFR measurements would also allow a more rigorous evaluation of continuous regression performance and of the consistency between the two physiological targets. Furthermore, our approach is validated on single coronary lesions (despite multiple minor lesions are already present in our dataset), since the value of FFR/iFR on diffuse diseases has still limited practical inferential use [63]. We also plan to employ specific synthetic data generation, such as [285], both as a form of augmentation and to share data in a privacy-preserving way.

Finally, we would like to highlight that our approach is thought for conventional angiography (using 2D+t data). It enables the estimation of blood fluid dynamics, while it is not able to reconstruct the geometry of arterial structures, for which FFR-CT procedure is required. We thus aim also to extend our approach with CT data to complement our findings, thereby offering a comprehensive approach to coronary artery disease assessment that can work on both conventional angiography via coronary catheterization and three-dimensional computational modeling using CT scans.

6.5 Technology Transfer and Intellectual Property

In parallel with the scientific validation reported in this article, part of the work on angiography-based functional assessment was pursued with a technology-transfer perspective, aimed at enabling translation into realistic cath-lab workflows. In this context, selected elements of the end-to-end pipeline for estimating functional indices (FFR/iFR) from routine coronary angiography have been the subject of intellectual property filings, including an Italian filing at UIBM and a subsequent European extension.[78, 77]

The filings are intended to cover system-level aspects that are directly relevant to deployment, such as the integration of video-based spatio-temporal modeling with procedure-aware preprocessing, inference orchestration, and workflow-compatible output generation. The scope is aligned with the objective of reducing time and cost compared to invasive pressure-wire measurements, while maintaining the clinical decision target of physiology-informed lesion assessment.

To remain consistent with ongoing IP processes, this thesis describes the contribution at

a scientific and architectural level, focusing on reproducible principles (e.g., spatio-temporal representation learning, calibration and evaluation protocols, robustness analyses), while avoiding claim-level parameters and implementation specifics that are not publicly available (e.g., proprietary thresholds, device-dependent integration details, and protected manufacturing or deployment configurations). This level of disclosure is sufficient to support academic scrutiny of the modeling choices and experimental evidence presented in the paper, without disclosing non-public elements that belong to the patent scope.

Where possible, the dissertation reports dataset splits, evaluation metrics, and ablation settings that enable independent comparison under similar conditions. Any clinical translation of the method should be considered within the constraints of local governance, site-specific acquisition protocols, and prospective validation requirements, and should be accompanied by appropriate safety, monitoring, and auditing procedures consistent with clinical use.

6.6 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 1200 hours of computation was performed on two NVIDIA Tesla T4 GPUs (thermal design power of 70 W), with total emissions estimated[2] to be 72.58 kg CO_2eq.

[2]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 7

Learning Joint Text and Visual Tokens in CLIP for Medical Image Analysis

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Concetto Spampinato, and Simone Palazzo. Learning joint text and visual tokens in clip for medical image analysis. In *Proceedings of the 2nd International Workshop on Multimedia Computing for Health and Medicine*, pages 45–51, 2025
- Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Ulas Bagci, Daniela Giordano, Concetto Spampinato, and Simone Palazzo. Conditional visuo-textual prompt learning for medical image analysis. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025

7.1 Overview

Recent advances at the intersection of computer vision and natural language processing have enabled large-scale vision-language models (VLMs) that learn aligned image-text representations from vast collections of weakly annotated web data [295, 162]. Among these approaches, contrastive pretraining has proven effective at producing transferable embeddings that can be reused for classification and retrieval through similarity-based inference, often under limited downstream supervision. This paradigm is appealing in medical imaging, where annotation is costly and cross-site robustness is essential. Yet, direct transfer from web-scale pretraining remains challenging: clinical concepts are fine-grained, terminology is specialized, and acquisition protocols differ substantially from natural images, leading to a

marked degradation of zero-shot performance.

A practical barrier to domain adaptation is that full fine-tuning of large backbones is frequently undesirable in clinical settings. Updating hundreds of millions of parameters requires substantial compute, increases the risk of overfitting when labeled data are scarce, and complicates reproducibility across institutions. These constraints have motivated parameter-efficient adaptation strategies that preserve pretrained weights while learning only a small set of task-specific parameters. In natural language processing, prompt learning emerged as a lightweight alternative to full fine-tuning: early work optimized discrete prompt templates or token selections [166, 333], while more recent continuous approaches learn a compact set of vectors directly in the embedding space [200, 194, 419]. Similar ideas have been adopted for VLMs, where learnable prompts can steer frozen encoders toward downstream tasks with minimal storage and training cost.

CLIP [295] demonstrated that contrastive alignment of images and text yields generalizable features, and subsequent work showed that scaling data and leveraging transformer backbones can further strengthen transfer [162, 373]. CLIP has become a backbone for a broad ecosystem of models, yet its utility in specialized clinical domains is limited by domain shift. Importantly, medical imaging introduces a dual mismatch: the visual stream must capture pathology-specific and procedure-specific appearances that are rare or absent in web images, and the text stream must represent specialist terminology and clinically meaningful label semantics. As a consequence, adapting only one modality (e.g., visual prompts alone or text prompts alone) can be insufficient, especially under heterogeneous acquisition conditions.

This chapter therefore focuses on joint prompt adaptation of both modalities in CLIP for coronary angiography classification. We first introduce Text and Visual Tokens Learnable in CLIP (TVTL-CLIP), a parameter-efficient framework that inserts distinct sets of soft prompts into the frozen CLIP text and image encoders (Fig. 7.1). The prompts steer the representation space toward task-relevant factors without modifying pretrained weights, enabling effective training under limited supervision. We evaluate TVTL-CLIP on two angiography benchmarks formulated as FFR 2-class and FFR 3-class tasks, and we show through systematic ablations that jointly adapting text and vision provides the most reliable trade-off across balanced accuracy, sensitivity/recall, specificity, and F1, while training orders of magnitude fewer parameters than full fine-tuning.

While TVTL-CLIP provides an efficient adaptation mechanism, clinical deployments often face additional variability beyond the task definition itself. The same label may appear differently across patients, devices, and acquisition protocols, and such factors can interact with patient variability and anatomy, reducing robustness under domain shift. To address this, we extend the framework to Conditional Text and Visual Tokens Learnable in CLIP

(CTVTL-CLIP), where the learnable tokens in both encoders are conditioned on the input instance via a lightweight support network. By allowing prompts to adapt at the sample level, CTVTL-CLIP targets improved robustness in challenging angiography settings, and its gains are most evident as task complexity and intra-class heterogeneity increase (e.g., multi-class severity grading), while maintaining the same constraint of parameter efficiency and compatibility with constrained training budgets.

Overall, TVTL-CLIP and its conditional extension CTVTL-CLIP operationalize the same design stance: rather than fine-tuning the entire VLM backbone, we adapt CLIP to medical tasks through a compact, trainable set of prompts in both modalities, and we further improve robustness by making this adaptation instance-aware when acquisition heterogeneity is a primary source of performance variability.

7.2 Method

We adapt a frozen CLIP backbone to domain-specific medical classification by learning a compact set of soft prompt tokens that steer both encoders while leaving all pretrained weights unchanged. The base formulation (TVTL-CLIP, Fig. 7.1) learns shared or class-specific prompts at the input of the CLIP text and vision transformers. The extended formulation (CTVTL-CLIP, Fig. 7.2) retains the same backbone and training objective, but conditions the prompts on the input instance through a lightweight support network, and can optionally strengthen visual adaptation through deep prompting across layers.

Let E_{vis} denote the CLIP vision encoder (ViT-B/32) and E_{text} the CLIP text transformer. Given an input image I, the vision encoder outputs a visual embedding $\mathbf{v}(I) \in \mathbb{R}^d$; given a class name c_k ($k \in \{1, \ldots, K\}$), the text encoder outputs a textual embedding $\mathbf{t}(c_k) \in \mathbb{R}^d$. Classification is performed through cosine similarities

$$s_k(I) \;=\; \frac{\langle \mathbf{v}(I), \mathbf{t}(c_k) \rangle}{\|\mathbf{v}(I)\| \, \|\mathbf{t}(c_k)\|}, \tag{7.1}$$

followed by a temperature-scaled softmax over classes.

TVTL-CLIP. TVTL-CLIP learns prompts jointly in the text and vision branches. For text prompting, for each class name c_k, we build a prompted sequence by prepending M learnable text tokens:

$$[\text{SOT}] \; \mathbf{p}^1 \ldots \mathbf{p}^M \; c_k \; [\text{EOT}], \tag{7.2}$$

where $\mathbf{p}^m \in \mathbb{R}^d$ are optimized while E_{text} remains frozen. We consider two variants: (i) shared context, where the same $\{\mathbf{p}^m\}_{m=1}^M$ is reused for all classes, and (ii) class-specific context

(CSC), where each class has its own $\{\mathbf{p}_k^m\}_{m=1}^M$. CSC scales linearly with the number of classes ($O(KM)$ parameters) and may become less convenient as K grows, so shared prompts are the default when scalability is a priority.

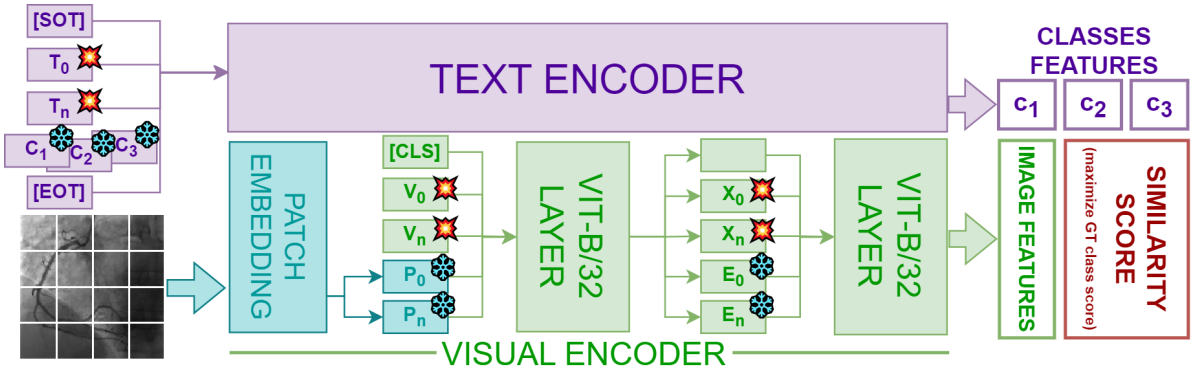


Figure 7.1: TVTL-CLIP architecture. The text encoder receives learnable tokens and class labels to produce a text embedding for each class. The image encoder (ViT-B/32) receives learnable visual tokens and patch embeddings to produce a visual embedding. Classification is performed by maximizing image-text similarity for the correct class

For visual prompting, let the ViT patch embedding produce a sequence of L patch tokens $\mathbf{x}_{1:L} \in \mathbb{R}^{d_v}$ (with $d_v = 768$ for ViT-B/32) and a class token $\mathbf{x}_{\text{cls}}$. TVTL-CLIP inserts Q learnable visual tokens $\mathbf{z}^{1:Q} \in \mathbb{R}^{d_v}$ between $\mathbf{x}_{\text{cls}}$ and the patch sequence, so that the input to the first transformer layer becomes

$$\left[\mathbf{x}_{\text{cls}}, \mathbf{z}^{1:Q}, \mathbf{x}_{1:L}\right]. \tag{7.3}$$

Only $\{\mathbf{z}^q\}_{q=1}^Q$ (and the text prompts) are trained; all CLIP backbone weights remain frozen.

Training objective. In supervised classification, we use the cosine similarities in Eq. (7.1) as logits against the K prompted class texts and optimize a temperature-scaled cross-entropy loss. This formulation is equivalent to the image-to-text direction of the CLIP objective when the batch is constructed so that each image is paired with its correct class text. For completeness, a CLIP-style symmetric contrastive loss over a mini-batch of size B can be written as

$$\mathcal{L}_{\text{TVTL}} = \frac{1}{B}\sum_{i=1}^{B}\Big[-\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} - \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}/\tau)}\Big], \tag{7.4}$$

where s_{ij} denotes the cosine similarity between image i and the paired text j within the batch, and τ is the temperature.

CTVTL-CLIP. CTVTL-CLIP extends TVTL-CLIP to improve robustness under acquisition heterogeneity by conditioning prompts on the input instance. As illustrated in Fig. 7.2,

a lightweight support network $S(\cdot)$ extracts an image-specific context vector $\mathbf{c}(I)$ used to modulate the learnable prompts, without unfreezing CLIP.

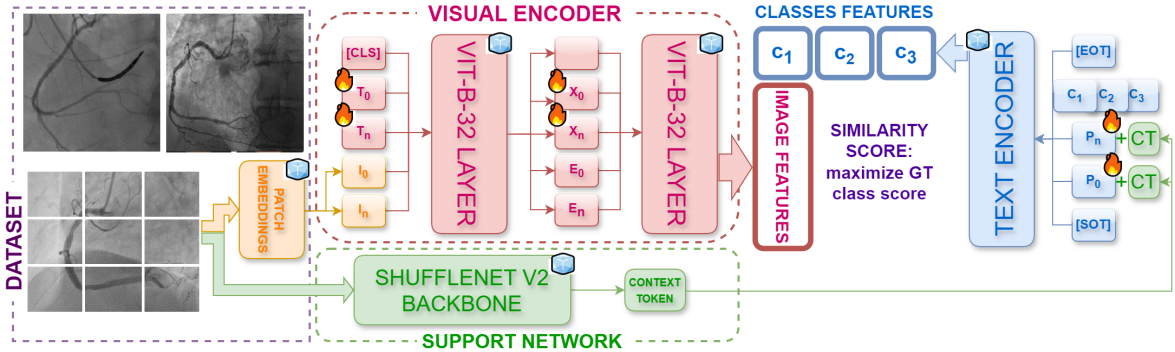


Figure 7.2: CTVTL-CLIP architecture. A support network produces an instance-specific context token that modulates the learnable prompts. The conditioned text prompts and class labels are encoded by the frozen text encoder, while the conditioned visual prompts and patch embeddings are encoded by the frozen vision encoder. The model predicts by image-text similarity, as in CLIP

We compute the context representation as

$$\mathbf{c}(I) = S(I), \tag{7.5}$$

where $S(\cdot)$ is lightweight. The context is projected into the text and vision prompt spaces through learned linear maps W_t and W_v:

$$\Delta_t(I) = W_t\,\mathbf{c}(I) \in \mathbb{R}^d, \qquad \Delta_v(I) = W_v\,\mathbf{c}(I) \in \mathbb{R}^{d_v}. \tag{7.6}$$

Conditional text prompts are formed by additive modulation of shared base prompts $\{\mathbf{p}^m\}_{m=1}^M$:

$$\tilde{\mathbf{p}}^m(I) = \mathbf{p}^m + g_t^m(\Delta_t(I))\,, \qquad m = 1, \ldots, M, \tag{7.7}$$

where $g_t^m(\cdot)$ can be (i) an identity map, adding the same $\Delta_t(I)$ to all prompt tokens, or (ii) a token-specific linear projection, enabling different modulations per m. The prompted class sequence becomes

$$[\text{SOT}]\ \tilde{\mathbf{p}}^1(I) \ldots \tilde{\mathbf{p}}^M(I)\ c_k\ [\text{EOT}], \tag{7.8}$$

and is encoded by the frozen E_{text}.

Conditional visual prompts are modulated analogously:

$$\tilde{\mathbf{z}}^q(I) = \mathbf{z}^q + g_v^q(\Delta_v(I))\,, \qquad q = 1, \ldots, Q, \tag{7.9}$$

and injected into the visual encoder as in Eq. (7.3), replacing $\mathbf{z}^{1:Q}$ with $\tilde{\mathbf{z}}^{1:Q}(I)$.

Deep visual prompting. To strengthen conditioning when domain shift primarily affects low-level and mid-level visual statistics, CTVTL-CLIP supports deep prompting. Let $\mathcal{V}_i(\cdot)$ denote the i-th transformer block of the vision encoder ($i = 1, \ldots, H$). Instead of injecting prompts only at the input, we maintain prompts per layer, $\{\mathbf{z}_i^{1:Q}\}$, and overwrite the corresponding prompt positions at each layer:

$$\mathbf{h}_i = \mathcal{V}_i\Big(\mathit{repQ}\big(\tilde{\mathbf{z}}_i^{1:Q}(I), \mathbf{h}_{i-1}\big)\Big), \tag{7.10}$$

where $\mathbf{h}_{i-1}$ is the previous layer output sequence and $\mathit{repQ}(\cdot)$ substitutes the Q prompt slots with the (optionally conditional) tokens $\tilde{\mathbf{z}}_i^{1:Q}(I)$. A lighter variant reuses the same Q prompts across all layers.

CTVTL-CLIP is trained with the same objective as Eq. (7.4), computed using the conditional prompts in Eqs. (7.8) and (7.9) (and, when enabled, the deep prompting mechanism in Eq. (7.10)). TVTL-CLIP is recovered as a special case by removing conditioning, i.e., $\mathbf{c}(I) = \mathbf{0}$, and using shallow prompting only.

7.3 Experimental Results

Dataset. We evaluate TVTL-CLIP and CTVTL-CLIP on a private, multi-center coronary angiography dataset retrospectively collected between January 2020 and January 2022. The cohort includes 389 patients diagnosed with CCS or ACS, acquired across five hospitals. Overall, the dataset comprises 778 angiographic acquisitions, typically including two projections per examination. Among the enrolled patients, 303 are men and 86 are women, with mean age 67.9 ± 9.61. The study was approved under IRB protocol number 0092163.

All patients underwent invasive physiological assessment of intermediate single-vessel coronary stenoses (30%-90% diameter stenosis by angiographic visual estimation) through FFR, iFR, or both. In particular, FFR measurements are available for 251 patients (64.5%), iFR for 228 patients (58.6%), and both indices for 90 patients (23.1%). For each angiographic acquisition, two expert cardiologists selected representative keyframes and annotated the major stenosis; disagreements were resolved by consensus. In this work, we intentionally adopt a 2D keyframe-based formulation: expert-selected frames concentrate the information around the most diagnostically relevant phase of contrast flow, reduce temporal redundancy and acquisition-dependent motion artifacts, and enable a controlled evaluation of prompt-based adaptation without introducing additional degrees of freedom from temporal modeling. This choice also reflects realistic deployment constraints, where fast triage or retrospective

reporting often relies on a small number of representative frames rather than full video sequences.

Acquisition settings vary across hospitals and devices: spatial resolution ranges from 512×512 to 1024×1024, and frame rates from 15 to 60 fps. Since our experiments operate on 2D keyframes, each frame is resized to 256×256 pixels before being fed to the model. To avoid information leakage across projections of the same patient, data splits are performed at the patient level; when single-view experiments are reported, only one randomly selected projection per patient is used in a given split, excluding the remaining views.

Label definition (FFR 2-class and FFR 3-class). We derive two classification benchmarks from the invasive indices by applying clinically motivated thresholds. In the binary setting (FFR 2-class), a stenosis is labeled as hemodynamically significant if FFR < 0.80 or iFR < 0.89, and non-significant otherwise. In the three-way setting (FFR 3-class), we define: (i) *low* severity when FFR ≥ 0.80 or iFR ≥ 0.89, (ii) *mild* severity when the measurement is below the previous threshold but remains above FFR 0.60 or iFR 0.69, and (iii) *high* severity when FFR < 0.60 or iFR < 0.69. When both FFR and iFR are available for the same case, the final label is assigned according to the more severe category (worst-case across indices), ensuring a conservative definition aligned with clinical risk stratification.

7.3.1 Training and Evaluation Procedure

TVTL-CLIP and CTVTL-CLIP are trained on the same gastroscopy dataset, with the goal of maximizing diagnostic performance while balancing sensitivity and specificity under pronounced class imbalance. In both methods, the CLIP backbone is kept frozen and only prompt-related parameters are optimized; CTVTL-CLIP additionally trains a lightweight *support network* used to condition the prompts on the input instance.

Optimization and efficiency. To enable efficient optimization of transformer-based encoders, training is performed with stochastic gradient descent (SGD) for 200 epochs, using a learning rate of 0.002. We employ 16-bit floating point training through CUDA Automatic Mixed Precision (AMP) to reduce memory footprint and accelerate training without visible degradation in accuracy. For CTVTL-CLIP, we additionally use a batch size of 128. (When not otherwise specified, the same optimization schedule is adopted across TVTL-CLIP experiments to ensure comparability.)

Loss function and class imbalance. TVTL-CLIP is trained by minimizing the symmetric contrastive objective defined in Eq. (5) (Section *Method*), computed over mini-batches of image-label pairs through all pairwise similarities. CTVTL-CLIP uses the same objective formulation, but adopts a *weighted* variant to mitigate class imbalance: we weight

the loss contributions according to the occurrence percentage of each class in the training set. This weighting is applied to the loss term referenced as Eq. (1) in the CTVTL-CLIP study, and is used by default in all conditional-prompt experiments.

Backbones and default configuration (CTVTL-CLIP). Unless otherwise stated, CTVTL-CLIP employs: (i) a ShuffleNet V2 backbone as *support network* to extract low-level contextual cues for conditioning, (ii) a ViT-B/32 model as the frozen CLIP visual encoder, and (iii) the CLIP text encoder as in [295]. Training jointly optimizes the learnable prompt parameters and the support network parameters.

Token ablations and prompt configurations (TVTL-CLIP). A key component of the TVTL-CLIP evaluation is the systematic exploration of token configurations for both modalities:

- *Text prompts.* We vary the number of learnable textual tokens from 0 up to 69 (the maximum feasible length for the pretrained text encoder once special tokens and the class-name tokens are accounted for). We evaluate both CSC prompting and non-CSC prompting
- *Visual prompts.* We vary the number of learnable visual tokens from 0 to 10. Visual tokens are inserted either (i) only at the input level (between the CLS token and the patch sequence) or (ii) through *deep visual prompting*, where visual tokens are injected at each of the 12 transformer layers of the ViT-B/32 visual encoder

From TVTL-CLIP to CTVTL-CLIP. CTVTL-CLIP retains the same training pipeline and frozen CLIP backbone, but extends TVTL-CLIP by conditioning prompts on the input instance through a support network. This modification is specifically introduced to improve robustness under heterogeneous acquisition conditions and intra-class variability, while keeping the optimization budget compatible with clinical constraints. We report results on two coronary angiography benchmarks, formulated as (i) a binary task (FFR 2-class) and (ii) a three-way severity task (FFR 3-class). For each benchmark, the evaluation follows two stages. First, we analyze prompt-design choices via a controlled grid search over textual and visual token configurations. Second, we benchmark the best prompt configuration against representative convolutional and transformer backbones trained either from scratch or by fine-tuning ImageNet-initialized weights. The same search space and selection protocol are used for TVTL-CLIP and CTVTL-CLIP, so that performance differences reflect instance-conditioned prompting rather than changes in the explored design space. Unless otherwise stated, Acc. denotes balanced accuracy.

Prompt-design ablations. Tables 7.1 and 7.3 summarize the token-configuration study for TVTL-CLIP, while Tables 7.2 and 7.4 report the corresponding results for CTVTL-CLIP.

Two consistent trends emerge across tasks and methods. First, adapting only one modality (text-only or vision-only prompting) is generally insufficient: the best settings consistently rely on jointly learning textual and visual prompts. Second, when visual tokens are used, deep visual prompting (tokens injected at multiple layers) is typically beneficial, supporting the view that layer-wise steering helps compensate for domain shift under a frozen visual encoder. For TVTL-CLIP, CSC can be toggled and yields limited, configuration-dependent effects; for CTVTL-CLIP, CSC is not a primary design axis in the available runs, hence values may repeat across CSC rows.

For the **FFR 2-class** task, TVTL-CLIP achieves its best configuration with multimodal prompting and deep visual token injection (Text Tok.=64, Vis. Tok.=5), reaching 64.8% Acc., 61.7% recall, 67.9% specificity, and 51.6 F1 (Table 7.1). On the same task, CTVTL-CLIP reaches 62.3% Acc. with 49.4% recall, 75.2% specificity, and 53.0 F1 (Table 7.2). Importantly, the best CTVTL performance is already attained by a compact prompt setting (Text Tok.=1, Vis. Tok.=1 with deep prompting), and the same operating point is also matched by larger token budgets. Overall, instance-conditioned prompting tends to shift the trade-off toward higher specificity (and higher F1 in the best setting), while remaining competitive in balanced accuracy.

For the **FFR 3-class** task, the ablations indicate a clearer gain from instance-conditioned prompting. TVTL-CLIP reaches 71.4% Acc. in its best configuration (Table 7.3), whereas CTVTL-CLIP improves to 73.6% Acc. with 55.1% recall and 62.7 F1 (Table 7.4). This multi-class setting is more sensitive to intra-class heterogeneity and ambiguous boundaries, and conditioning prompts on the input instance appears to support more robust severity grading.

Benchmark against SOTA backbones. After selecting the best configuration per method and task from the ablation study, we compare against a set of CNN and transformer backbones. Tables 7.5 and 7.6 report results for (i) training from scratch and (ii) fine-tuning from ImageNet initialization, together with TVTL-CLIP and CTVTL-CLIP. In both tasks, ImageNet initialization does not consistently improve performance, suggesting that generic natural-image pretraining only partially transfers to angiography and that optimization can be sensitive to the chosen backbone and training regime.

On **FFR 2-class** (Table 7.5), TVTL-CLIP attains the highest balanced accuracy among the compared methods (64.8%) while training only 0.08M parameters. CTVTL-CLIP is competitive in balanced accuracy (62.3%) and achieves a higher specificity operating point (75.2%) under the same parameter budget, reflecting a different trade-off that can be preferable when false positives are more costly. On **FFR 3-class** (Table 7.6), CTVTL-CLIP achieves the strongest overall results in this benchmark, reaching 73.6% balanced accuracy and 62.7 F1

Table 7.1: Performance comparison with different combinations of textual and visual tokens on the FFR 2-class TVTL-CLIP. *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer.

Text Tok.	CSC	Vis. Tok.	Deep Vis. Tok.	Acc.	Recall	Spec.	F1 Score
0		0		44.6	17.3	71.9	21.7
0		1	✗	45.9	22.2	69.6	26.4
0		1	✓	48.8	30.9	66.7	34.6
0		5	✗	53.9	37.0	70.7	38.7
0		5	✓	51.4	34.6	68.2	35.8
1	✗	0		44.6	17.3	71.9	21.7
1	✓	0		44.6	17.3	71.9	21.7
1	✗	1	✗	46.6	22.2	71.1	25.2
1	✗	1	✓	44.2	16.0	72.4	19.2
1	✓	1	✗	46.6	22.2	71.1	25.2
1	✓	1	✓	44.2	16.0	72.4	19.2
1	✗	5	✗	51.2	35.8	66.7	36.5
1	✗	5	✓	49.4	32.1	66.7	33.8
1	✓	5	✗	51.2	35.8	66.7	36.5
1	✓	5	✓	49.4	32.1	66.7	33.8
64	✗	0		56.9	43.2	70.7	43.0
64	✓	0		56.9	43.2	70.7	43.0
64	✗	1	✗	59.3	50.6	67.9	46.9
64	✗	1	✓	55.8	39.5	72.1	39.7
64	✓	1	✗	59.3	50.6	67.9	46.9
64	✓	1	✓	55.8	39.5	72.1	39.7
64	✗	5	✗	64.8	61.7	67.9	51.6
64	✗	5	✓	61.3	58.0	64.6	47.8
64	✓	5	✗	61.3	58.0	64.6	47.8
64	✓	5	✓	**64.8**	**61.7**	**67.9**	**51.6**

Table 7.2: Performance comparison with different combinations of textual and visual tokens on the FFR 2-class CTVTL-CLIP (ShuffleNet V2 support). *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer.

Text Tok.	CSC	Vis. Tok.	Deep Vis. Tok.	Acc.	Recall	Spec.	F1 Score
0		0		45.9	17.3	74.4	21.4
0		1	✗	44.9	14.8	75.0	18.2
0		1	✓	52.4	28.4	76.4	32.4
0		5	✗	52.7	29.6	75.8	33.3
0		5	✓	53.1	32.1	74.0	35.3
1	✗	0		49.3	23.5	75.0	28.6
1	✓	0		49.3	23.5	75.0	28.6
1	✗	1	✗	58.2	37.0	79.3	40.5
1	✗	1	✓	62.3	49.4	75.2	53.0
1	✓	1	✗	58.2	37.0	79.3	40.5
1	✓	1	✓	53.7	33.3	74.0	36.8
1	✗	5	✗	50.9	28.4	73.3	32.3
1	✗	5	✓	47.8	21.0	74.6	24.6
1	✓	5	✗	50.9	28.4	73.3	32.3
1	✓	5	✓	47.8	21.0	74.6	24.6
64	✗	0		55.4	39.5	71.1	39.8
64	✓	0		55.4	39.5	71.1	39.8
64	✗	1	✗	54.7	38.3	71.1	38.8
64	✗	1	✓	52.1	33.3	71.0	34.9
64	✓	1	✗	54.7	38.3	71.1	38.8
64	✓	1	✓	52.1	33.3	71.0	34.9
64	✗	5	✗	61.0	54.3	67.8	53.7
64	✗	5	✓	53.7	33.3	74.0	36.8
64	✓	5	✗	61.0	54.3	67.8	53.7
64	✓	5	✓	**62.3**	**49.4**	**75.2**	**53.0**

Table 7.3: Performance comparison with different combinations of textual and visual tokens on the FFR 3-class TVTL-CLIP. *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer.

Text Tok.	CSC	Vis. Tok.	Deep Vis. Tok.	Acc.	Recall	Spec.	F1 Score
0		0		55.8	15.7	95.8	21.6
0		1	✗	55.8	14.0	96.4	20.7
0		1	✓	54.0	11.0	96.8	16.6
0		5	✗	61.1	27.1	96.9	34.3
0		5	✓	62.0	27.9	96.2	35.2
1	✗	0		55.8	15.7	95.8	21.6
1	✓	0		55.8	15.7	95.8	21.6
1	✗	1	✗	56.9	16.5	96.2	22.1
1	✗	1	✓	59.1	22.3	95.8	28.3
1	✓	1	✗	56.9	16.5	96.2	22.1
1	✓	1	✓	59.1	22.3	95.8	28.3
1	✗	5	✗	66.7	35.5	95.8	43.9
1	✗	5	✓	69.9	43.5	94.6	51.2
1	✓	5	✗	66.7	35.5	95.8	43.9
1	✓	5	✓	69.9	43.5	94.6	51.2
64	✗	0		67.1	35.9	96.9	44.6
64	✓	0		67.1	35.9	96.9	44.6
64	✗	1	✗	70.8	45.0	96.6	53.4
64	✗	1	✓	68.0	38.4	96.4	47.7
64	✓	1	✗	70.8	45.0	96.6	53.4
64	✓	1	✓	68.0	38.4	96.4	47.7
64	✗	5	✗	71.4	43.5	97.0	53.2
64	✗	5	✓	62.0	27.9	96.2	35.2
64	✓	5	✗	62.0	27.9	96.2	35.2
64	✓	5	✓	**71.4**	**43.5**	**97.0**	**53.2**

Table 7.4: Performance comparison with different combinations of textual and visual tokens on the FFR 3-class CTVTL-CLIP (ShuffleNet V2 support). *Text Tok.* indicates the number of textual tokens used, *CSC* indicates whether a class-specific context was used, *Vis. Tok.* indicates the number of visual tokens added, *Deep Vis. Tok.* indicates whether visual tokens were added at each visual transformation layer.

Text Tok.	CSC	Vis. Tok.	Deep Vis. Tok.	Acc.	Recall	Spec.	F1 Score
0		0		67.4	39.6	95.3	47.9
0		1	✗	69.4	44.9	94.3	52.3
0		1	✓	67.3	39.0	95.5	46.9
0		5	✗	72.3	52.4	95.3	59.7
0		5	✓	70.1	46.9	94.0	54.8
1	✗	0		67.0	38.3	95.3	46.3
1	✓	0		67.0	38.3	95.3	46.3
1	✗	1	✗	67.1	38.6	95.5	47.2
1	✗	1	✓	66.9	38.0	95.3	46.4
1	✓	1	✗	67.1	38.6	95.5	47.2
1	✓	1	✓	66.9	38.0	95.3	46.4
1	✗	5	✗	70.1	46.9	94.0	54.8
1	✗	5	✓	69.1	44.4	93.7	51.7
1	✓	5	✗	70.1	46.9	94.0	54.8
1	✓	5	✓	69.1	44.4	93.7	51.7
64	✗	0		66.6	37.4	95.8	45.2
64	✓	0		66.6	37.4	95.8	45.2
64	✗	1	✗	66.9	38.1	95.0	46.6
64	✗	1	✓	67.5	39.6	95.8	48.2
64	✓	1	✗	66.9	38.1	95.0	46.6
64	✓	1	✓	67.5	39.6	95.8	48.2
64	✗	5	✗	69.9	46.2	93.7	54.1
64	✗	5	✓	70.1	46.9	94.0	54.8
64	✓	5	✗	69.9	46.2	93.7	54.1
64	✓	5	✓	**73.6**	**55.1**	**95.7**	**62.7**

Table 7.5: Comparison with SOTA architectures on the FFR 2-class benchmark. "# params" reports trainable weights in millions. For each backbone we list results when trained from scratch and when finetuned from an ImageNet initialization (grey rows). All metrics are percentages; Acc. reports balanced accuracy. TVTL-CLIP and CTVTL-CLIP are shown using their best configuration on this benchmark.

Model	# params ($\times 10^6$)	Acc.	Recall	Spec.	F1 Score
DenseNet-121	6.96	60.0	84.3	35.8	61.6
↪ Pre-trained		58.8	84.3	33.3	60.4
DenseNet-161	26.49	64.2	82.6	45.7	**65.4**
↪ Pre-trained		56.3	79.3	33.3	57.1
DenseNet-169	12.50	63.8	84.3	43.2	65.3
↪ Pre-trained		56.6	84.3	28.9	60.2
DenseNet-201	18.11	63.4	84.3	42.6	65.1
↪ Pre-trained		57.4	**84.6**	30.2	61.0
RegNet	105.31	60.9	73.6	48.1	62.5
↪ Pre-trained		56.8	75.2	38.3	58.0
ResNet-18	11.18	55.8	61.2	50.4	56.5
↪ Pre-trained		55.6	67.8	43.4	56.5
ResNet-50	23.52	57.0	67.8	46.3	57.3
↪ Pre-trained		55.8	67.8	43.8	56.6
ResNet-101	42.52	55.4	62.0	48.8	56.1
↪ Pre-trained		55.8	62.0	49.6	56.5
SqueezeNet	0.73	56.1	57.0	55.1	55.7
↪ Pre-trained		53.2	59.5	46.9	52.1
Swin Transformer v2	86.91	54.6	55.4	53.7	54.6
↪ Pre-trained		55.8	66.1	45.5	56.3
ViT-B-16	85.80	55.1	59.5	50.6	55.0
↪ Pre-trained		55.8	66.1	45.5	56.3
ViT-B-32	87.46	56.1	68.6	43.6	57.0
↪ Pre-trained		55.8	69.4	42.1	56.8
ViT-L-16	303.52	52.2	52.9	51.4	51.6
↪ Pre-trained		55.8	66.1	45.5	56.3
ViT-L-32	305.52	54.5	55.4	53.7	54.0
↪ Pre-trained		56.5	66.1	46.9	56.5
TVTL-CLIP	**0.08**	**64.8**	61.7	67.9	51.6
CTVTL-CLIP	**0.08**	62.3	49.4	**75.2**	53.0

Table 7.6: Comparison with SOTA architectures on the FFR 3-class benchmark. "# params" reports trainable weights in millions. For each backbone we list results when trained from scratch and when finetuned from an ImageNet initialization (grey rows). All metrics are percentages; Acc. reports balanced accuracy. TVTL-CLIP and CTVTL-CLIP are shown using their best configuration on this benchmark.

Model	# params ($\times 10^6$)	Acc.	Recall	Spec.	F1 Score
DenseNet-121	6.96	71.7	46.5	95.7	48.6
↪ Pre-trained		70.7	43.9	95.3	45.9
DenseNet-161	26.49	71.8	46.0	95.5	48.0
↪ Pre-trained		68.3	37.7	95.5	38.9
DenseNet-169	12.50	71.4	44.7	95.5	46.8
↪ Pre-trained		68.0	38.6	95.1	38.3
DenseNet-201	18.11	71.1	44.7	95.0	46.2
↪ Pre-trained		68.0	38.6	95.1	38.3
RegNet	105.31	69.5	40.4	95.1	41.8
↪ Pre-trained		69.2	40.0	95.3	41.3
ResNet-18	11.18	67.7	36.8	95.5	37.8
↪ Pre-trained		66.8	35.5	95.1	36.4
ResNet-50	23.52	67.4	36.8	95.1	37.4
↪ Pre-trained		66.8	35.5	95.1	36.4
ResNet-101	42.52	66.8	35.5	95.1	36.4
↪ Pre-trained		66.8	35.5	95.1	36.4
SqueezeNet	0.73	66.0	34.2	94.8	34.7
↪ Pre-trained		68.2	38.6	95.1	39.1
Swin Transformer v2	86.91	66.0	34.2	94.8	34.7
↪ Pre-trained		68.0	38.6	95.1	38.3
ViT-B-16	85.80	68.0	38.6	95.1	38.3
↪ Pre-trained		68.0	38.6	95.1	38.3
ViT-B-32	87.46	68.5	39.4	95.1	39.7
↪ Pre-trained		68.0	38.6	95.1	38.3
ViT-L-16	303.52	66.8	35.5	95.1	36.4
↪ Pre-trained		66.8	**100.0**	33.5	50.2
ViT-L-32	305.52	66.0	34.2	94.8	34.7
↪ Pre-trained		68.0	38.6	95.1	38.3
TVTL-CLIP	**0.08**	71.4	43.5	**97.0**	53.2
CTVTL-CLIP	**0.08**	**73.6**	55.1	95.7	**62.7**

while still training only 0.08M parameters. Compared to conventional fine-tuning baselines that require tens to hundreds of millions of trainable weights, multimodal prompt learning achieves comparable or better performance with a reduction of roughly two to three orders of magnitude in trainable parameters, and instance-conditioned prompting provides the clearest gains as task complexity and intra-class variability increase.

7.3.2 Ablation Studies

Since CTVTL-CLIP is conceived as an extension of TVTL-CLIP, we present a unified ablation analysis that (i) retains the original TVTL-CLIP axes (text-only, visual-only, and CLIP backbone), and (ii) incorporates the additional components introduced by CTVTL-CLIP, namely the instance-conditioned context token (CT) and the support network. For clarity and reproducibility, we include all non-duplicated tables across the two versions. When a table (or subset of rows) was reported as commented-out in the CTVTL-CLIP paper and the corresponding results are consistent with the TVTL-CLIP runs (same dataset split, metrics, and matching values), we preserve that status as-is to document what was reproduced without introducing redundant duplicates.

Text-only ablation

To isolate the effect of textual prompting, we disable visual prompting and vary the number of learnable text tokens, testing both shared and CSC when applicable. This protocol quantifies how far language-side adaptation alone can compensate for the domain shift of coronary angiography when the CLIP visual encoder is kept frozen.

On the FFR 2-class benchmark (Tab. 7.7), text-only prompting yields modest and non-monotonic gains compared to the no-prompt baseline. Across token counts, TVTL exhibits small improvements that do not scale reliably with prompt length, and CSC provides at most incremental benefits. CTVTL shows a similar operating regime, where instance-conditioned prompting can improve some configurations but does not transform the overall picture when visual prompts are absent. Overall, these results indicate that language-side adaptation alone is insufficient to recover the task-relevant visual separability required by angiography: meaningful gains require directly steering the visual representation, rather than relying on textual prompts as a proxy.

A consistent pattern emerges on the FFR 3-class benchmark (Tab. 7.8). While both methods improve over the baseline, the dependence on the number of text tokens remains weak and sometimes inconsistent across settings. When CSC is enabled for TVTL, the gains are limited and do not systematically translate into higher balanced accuracy or

Table 7.7: Text-only ablation study on the FFR 2-class benchmark. Performance is reported when no visual prompts are used. "*Text Tok.*" denotes the number of learnable textual prompt tokens, while "*CSC*" marks whether class-specific context is applied or not. TVTL and CTVTL are reported side-by-side; since CSC is not varied in CTVTL, the same values are repeated across CSC rows. All metrics are expressed in %.

Text Tok.	CSC	TVTL				CTVTL			
		Acc.	Recall	Spec.	F1 Score	Acc.	Recall	Spec.	F1 Score
0	-	44.6	17.3	71.9	21.7	44.6	17.3	71.9	21.7
1	✗	57.2	18.5	95.9	29.7	59.0	27.2	90.9	38.6
1	✓	57.2	18.5	95.9	29.7	59.0	27.2	90.9	38.6
2	✗	57.2	30.9	83.5	39.7	58.0	28.4	87.6	38.7
2	✓	57.2	30.9	83.5	39.7	58.0	28.4	87.6	38.7
4	✗	57.6	35.8	79.3	43.0	61.7	38.3	85.1	47.7
4	✓	57.6	35.8	79.3	43.0	61.7	38.3	85.1	47.7
8	✗	57.8	37.0	78.5	43.8	51.5	56.8	46.3	47.9
8	✓	59.4	42.0	76.9	47.6	51.5	56.8	46.3	47.9
16	✗	57.2	38.3	76.0	44.0	61.1	44.4	77.7	50.0
16	✓	57.2	38.3	76.0	44.0	61.1	44.4	77.7	50.0
32	✗	58.4	43.2	73.6	47.3	61.3	53.1	69.4	53.4
32	✓	58.4	43.2	73.6	47.3	61.3	53.1	69.4	53.4
64	✗	58.4	38.3	78.5	44.9	61.8	51.2	72.4	47.6
64	✓	58.4	38.3	78.5	44.9	61.8	51.2	72.4	47.6
Best overall	-	**64.8**	**61.7**	**67.9**	**51.6**	**62.3**	**49.4**	**75.2**	**53.0**

F1. In summary, across both binary and three-class severity grading, textual prompting alone provides only partial adaptation and does not close the gap to multimodal prompt configurations, reinforcing the need to adapt the visual branch under frozen encoders.

Visual-only ablation

To isolate the contribution of visual prompting, we disable learnable textual tokens and vary the number of visual tokens (0-10), comparing shallow injection (tokens added only at the visual input) against deep prompting (tokens injected at each transformer block). This analysis tests whether steering the visual pathway alone can restore discriminative performance under the angiography domain shift.

The visual-only results (Tabs. 7.9 and 7.10) show a markedly stronger effect than the corresponding text-only setting. Shallow prompting yields limited or unstable improvements, whereas deep prompting consistently increases balanced accuracy and F1 across token counts, suggesting that layer-wise steering is needed to effectively modulate frozen CLIP representations. In practice, a small number of deep tokens already achieves most of the attainable gains, while larger token budgets provide diminishing returns and can introduce variability across configurations.

Across both tasks, performance tends to saturate once a small set of deep visual tokens is

Table 7.8: Text-only ablation study on the FFR 3-class benchmark. Performance is reported when no visual prompts are used. "*Text Tok.*" denotes the number of learnable textual prompt tokens, while "*CSC*" marks whether class-specific context is applied or not. TVTL and CTVTL are reported side-by-side; since CSC is not varied in CTVTL, the same values are repeated across CSC rows. All metrics are expressed in %.

Text Tok.	CSC	TVTL				CTVTL			
		Acc.	Recall	Spec.	F1 Score	Acc.	Recall	Spec.	F1 Score
0	-	52.7	39.0	66.4	17.0	52.2	37.4	67.0	16.6
1	✗	64.4	32.2	96.7	40.7	64.7	32.6	96.7	40.9
1	✓	64.4	32.2	96.7	40.7	64.7	32.6	96.7	40.9
2	✗	67.3	39.2	95.5	46.8	67.7	40.9	94.6	46.4
2	✓	67.3	39.2	95.5	46.8	67.7	40.9	94.6	46.4
4	✗	65.4	35.1	95.8	40.7	67.5	39.6	95.4	43.9
4	✓	65.4	35.1	95.8	40.7	67.5	39.6	95.4	43.9
8	✗	66.8	38.2	95.5	43.4	66.3	37.5	95.2	41.1
8	✓	66.8	38.2	95.5	43.4	66.3	37.5	95.2	41.1
16	✗	68.6	42.9	94.4	45.9	67.1	38.2	96.0	44.1
16	✓	68.2	42.5	93.9	44.5	67.1	38.2	96.0	44.1
32	✗	66.7	37.6	95.9	44.7	64.7	34.0	95.3	38.9
32	✓	65.9	36.1	95.8	41.8	64.7	34.0	95.3	38.9
64	✗	66.9	39.2	94.6	41.9	68.6	41.1	96.1	43.2
64	✓	67.2	39.2	95.3	44.3	68.6	41.1	96.1	43.2
Best overall	-	**71.4**	**43.5**	**97.0**	**53.2**	**73.6**	**55.1**	**95.7**	**62.7**

Table 7.9: Results using only learnable visual tokens on the FFR 2-class benchmark (no textual prompts). (*Vis. Tok.* number of visual tokens, *Deep Vis. Tok.* w/wo layer-wise visual tokens). TVTL and CTVTL are reported side-by-side.

Vis. Tok.	**Deep Vis. Tok.**	**TVTL**				**CTVTL**			
		Acc. Bal.	**Sens.**	**Spec.**	**F1 Score**	**Acc. Bal.**	**Sens.**	**Spec.**	**F1 Score**
0		44.6	17.3	71.9	21.7	44.6	17.3	71.9	21.7
1	✗	57.0	24.7	89.3	35.1	57.6	18.5	96.7	30.0
1	✓	59.0	51.9	66.1	51.2	61.5	56.1	66.9	56.0
3	✗	52.8	22.2	83.5	30.3	57.2	25.9	88.4	36.2
3	✓	65.6	54.3	76.9	57.5	58.8	50.6	66.9	50.6
5	✗	57.0	27.2	86.8	37.0	52.7	29.6	75.8	33.3
5	✓	60.4	48.1	72.7	51.0	53.1	32.1	74.0	35.3
6	✗	53.2	23.5	82.9	32.5	57.2	25.9	88.4	36.2
6	✓	61.7	54.3	69.0	53.7	60.8	50.6	71.1	52.2
7	✗	52.2	21.0	83.5	29.4	60.1	30.9	89.3	42.0
7	✓	60.9	55.6	66.1	53.6	65.0	53.1	76.9	56.6
8	✗	55.8	28.4	83.1	37.8	57.8	27.2	88.4	37.6
8	✓	60.0	56.8	63.2	52.6	64.8	56.8	72.7	57.5
9	✗	56.5	25.9	87.2	36.4	58.6	27.2	90.1	38.3
9	✓	61.7	56.8	66.5	53.6	61.3	45.7	76.9	50.7
10	✗	54.5	23.5	85.5	33.0	60.5	30.9	90.1	42.4
10	✓	61.3	54.3	68.2	53.2	59.6	70.4	48.8	57.0
Best overall		**64.8**	**61.7**	**67.9**	**51.6**	**62.3**	**49.4**	**75.2**	**53.0**

Table 7.10: Results using only learnable visual tokens on the FFR 3-class benchmark (no textual prompts). (*Vis. Tok.* number of visual tokens, *Deep Vis. Tok.* w/wo layer-wise visual tokens). TVTL and CTVTL are reported side-by-side.

Vis. Tok.	Deep Vis. Tok.	TVTL				CTVTL			
		Acc. Bal.	Sens.	Spec.	F1 Score	Acc. Bal.	Sens.	Spec.	F1 Score
0		52.7	39.0	66.4	17.0	52.2	37.4	67.0	16.6
1	✗	50.1	0.4	99.8	0.8	50.4	1.2	99.6	2.3
1	✓	70.2	46.3	94.1	47.4	72.8	50.8	94.9	51.5
3	✗	52.7	7.4	97.9	10.3	52.8	7.0	98.6	10.6
3	✓	70.1	45.7	94.5	47.9	71.2	47.2	95.1	49.1
5	✗	55.8	14.5	97.1	20.7	55.7	14.3	97.1	20.7
5	✓	73.9	52.2	95.6	54.4	71.6	47.5	95.7	62.7
6	✗	54.8	12.1	97.5	17.3	57.0	17.1	96.9	21.7
6	✓	70.6	46.1	95.1	51.1	72.6	49.9	95.2	51.3
7	✗	55.4	12.8	98.0	18.2	59.2	22.3	96.1	28.6
7	✓	71.0	47.3	94.7	51.6	72.7	49.9	95.4	52.0
8	✗	56.2	12.5	99.8	19.5	60.9	25.4	96.3	32.4
8	✓	70.8	47.4	94.2	51.2	72.4	49.3	95.5	51.0
9	✗	56.8	14.1	99.6	21.0	59.7	22.9	96.5	30.3
9	✓	71.1	47.5	94.7	51.7	74.1	53.2	95.0	53.6
10	✗	56.1	13.6	98.5	19.3	58.6	19.2	98.0	28.8
10	✓	71.1	47.8	94.5	47.9	71.5	48.1	94.9	49.1
Best overall		**71.4**	**43.5**	**97.0**	**53.2**	**73.6**	**55.1**	**95.7**	**62.7**

used, with additional tokens offering limited improvements and occasional instability. Importantly, visual-only prompting does not replace the multimodal design: the best configurations in the token-combination study remain stronger than visual-only results, indicating that text and vision prompts act as complementary mechanisms rather than interchangeable ones.

CLIP backbone ablation

We next assess the impact of the frozen CLIP visual backbone while keeping the prompt configuration fixed (Text Tok.=64, CSC=True, Vis. Tok.=5, Deep Vis. Tok.=Yes). This experiment isolates how the inductive biases and representational quality of the underlying visual encoder interact with prompt-based adaptation under a constant training protocol.

On the FFR 2-class benchmark (Tab. 7.11), different CLIP backbones lead to sizable variation in balanced accuracy and in the sensitivity-specificity trade-off under the same prompt setup. ViT-based encoders provide a competitive operating point, while ResNet-based variants span a broader range of behaviors. In this setting, CTVTL often shifts the trade-off toward higher recall or improved F1 under several backbones, indicating that the instance-conditioned signal can refine decision boundaries even when the backbone is frozen.

A similar conclusion holds on the FFR 3-class benchmark (Tab. 7.12), where backbone choice has an even stronger impact due to the increased difficulty of multi-class severity grading. Under the same fixed prompting configuration, CTVTL generally provides more robust multi-class behavior, often improving balanced accuracy and/or F1 while maintaining

Table 7.11: Ablation on different CLIP visual backbones on the FFR 2-class benchmark, using the fixed prompt setup of TVTL (Text Tok.=64, CSC=True, Vis. Tok.=5, Deep Vis. Tok.=Yes). For CTVTL, the same token setup is used (CSC is fixed by design in the available runs).

Visual Encoder Model	TVTL				CTVTL			
	Acc.	Recall	Spec.	F1 Score	Acc.	Recall	Spec.	F1 Score
ResNet-50	53.9	33.3	74.4	38.8	59.2	48.1	70.2	50.0
ResNet-50x4	58.4	38.3	78.5	44.9	58.8	33.3	84.3	42.5
ResNet-50x16	55.5	30.9	80.2	38.5	55.7	34.6	76.9	40.9
ResNet-50x64	55.5	33.3	77.7	40.0	54.9	42.0	67.8	44.2
ResNet-101	56.5	32.1	81.0	40.0	55.5	25.9	85.1	35.0
ViT-B-16	54.0	42.0	66.1	43.6	53.6	46.9	60.3	45.5
ViT-B-32 (Ours)	**64.8**	**61.7**	**67.9**	**51.6**	**62.3**	**49.4**	**75.2**	**53.0**

strong specificity. These results suggest that instance-conditioned prompting is most beneficial when class boundaries are less separable and the model must resolve finer-grained distinctions.

Table 7.12: Ablation on different CLIP visual backbones on the FFR 3-class benchmark, using the fixed prompt setup of TVTL (Text Tok.=64, CSC=True, Vis. Tok.=5, Deep Vis. Tok.=Yes). For CTVTL, the same token setup is used (CSC is fixed by design in the available runs).

Visual Encoder Model	TVTL				CTVTL			
	Acc.	Recall	Spec.	F1 Score	Acc.	Recall	Spec.	F1 Score
ResNet-50	62.5	30.1	94.9	33.7	64.1	31.6	96.7	39.9
ResNet-50x4	60.1	23.1	97.0	30.8	63.9	31.1	96.6	38.5
ResNet-50x16	65.7	36.3	95.0	40.3	66.0	36.9	95.1	40.8
ResNet-50x64	58.5	19.4	97.6	27.7	54.0	10.7	97.3	15.3
ResNet-101	60.2	24.4	96.1	32.4	66.5	36.7	96.3	43.5
ViT-B-16	70.7	47.0	94.4	46.9	70.8	45.8	95.8	49.7
ViT-B-32 (Ours)	**71.4**	**43.5**	**97.0**	**53.2**	**73.6**	**55.1**	**95.7**	**62.7**

Overall, the ablation study supports three consistent takeaways across the FFR 2-class and FFR 3-class settings. First, visual prompting is the main driver of adaptation under a frozen CLIP encoder, whereas text-only prompting provides limited gains; nevertheless, combining text and vision prompts remains beneficial, supporting a multimodal prompt design rather than single-branch adaptation. Second, deep visual prompting is consistently more effective than shallow injection, indicating that layer-wise steering is required to obtain stable improvements under angiography domain shift. Third, CTVTL-CLIP improves over TVTL-CLIP by injecting CT through the support network, yielding a more favorable accuracy-robustness trade-off, with the clearest gains emerging in the three-class severity task while preserving high specificity.

7.4 Discussion

This chapter consolidates our investigation on parameter-efficient adaptation of large-scale vision-language models to coronary angiography classification through prompt learning. Starting from TVTL-CLIP, we showed that a frozen CLIP backbone can be specialized to two clinically relevant tasks, FFR 2-class significance detection and FFR 3-class severity grading, by optimizing a compact set of learnable tokens in both the vision and text encoders. Despite training fewer than 0.1M parameters, TVTL-CLIP matches or surpasses strong convolutional and transformer baselines on both benchmarks, supporting the view that joint visual-textual prompting is an effective mechanism to mitigate domain shift in angiography under limited supervision.

CTVTL-CLIP extends this formulation by introducing instance-conditioned prompting through a lightweight support network that produces a CT. This design is motivated by the heterogeneous nature of coronary angiography, where intra-class variability, differences in acquisition protocols and devices, and patient-specific factors can lead to markedly different visual manifestations for the same functional label. By conditioning the prompting mechanism on the input instance, CTVTL-CLIP improves robustness to such variability while retaining the key advantages of the TVTL-CLIP paradigm, frozen foundation backbones, small trainable parameter budgets, and similarity-based inference in a shared embedding space. Consistently with the ablation study, the conditional formulation can also reach strong operating points with compact prompt settings, indicating that robustness gains do not necessarily require large token budgets.

Across the unified ablation analysis, three consistent patterns emerge. First, visual prompting is the main driver of domain adaptation: injecting learnable visual tokens substantially improves sensitivity and balanced performance, and deep visual prompting is particularly effective, suggesting that steering intermediate visual representations is important when the pretrained vision encoder has limited exposure to angiographic appearance statistics. Second, textual prompting provides complementary gains by refining class discrimination in the shared embedding space, but its contribution is most reliable when paired with visual prompting rather than used in isolation. Third, jointly learning visual and textual prompts yields the most stable trade-off across balanced accuracy, sensitivity/recall, specificity, and F1, confirming that transfer limitations in medical VLM deployment affect both modalities and are best addressed jointly.

The comparison against representative CNN and transformer backbones further highlights a practical point for deployment: ImageNet initialization does not consistently improve performance on coronary angiography, and in several cases training from scratch is competitive

or superior. In contrast, TVTL-CLIP and CTVTL-CLIP provide strong results while updating only 0.08M parameters, offering an attractive alternative to full fine-tuning when compute, data volume, and reproducibility constraints are central. In addition, the keyframe-based 2D formulation adopted in this chapter provides a controlled setting to isolate prompt effects, while reflecting realistic workflows where a small number of diagnostically relevant frames is often used for fast triage or retrospective reporting.

A limitation of the proposed prompt-based formulation is that the visual encoder remains frozen and is inherited from a generic vision-language model. This choice was deliberate: the medical datasets considered in this chapter are relatively small compared with the scale usually required to train or fully fine-tune transformer backbones reliably, and updating the entire encoder could increase overfitting risk, reduce reproducibility, and weaken transfer under domain shift. Freezing the backbone and optimizing only a compact set of visual and textual tokens therefore provides a parameter-efficient adaptation strategy that is better aligned with limited supervision and low computational cost. At the same time, a frozen generic encoder may fail to capture fine pathology-specific textures or subtle appearance patterns that are underrepresented in generic pretraining data. If sufficiently large and well-annotated medical datasets were available, full fine-tuning or domain-specific pretraining could improve the representation of such details. Future work should therefore compare the current frozen-backbone strategy with intermediate adaptation schemes, including partial unfreezing, LoRA/adapters, domain-specific self-supervised pretraining, and staged training protocols in which prompt tuning is followed by carefully regularized backbone adaptation.

From a deployment perspective, these findings suggest an adaptation recipe compatible with common clinical constraints: limited annotated data, limited compute budgets, and the need for reproducibility across sites. The parameter-light nature of TVTL-CLIP/CTVTL-CLIP reduces the risk of overfitting and simplifies model distribution compared to full fine-tuning, while preserving the representational capacity inherited from web-scale contrastive pretraining. Moreover, CT-based conditioning provides a principled mechanism to cope with heterogeneous acquisition conditions without enlarging the trainable backbone. Notably, its benefits are more pronounced in the multi-class severity setting, where intra-class heterogeneity and ambiguous decision boundaries are expected to be stronger.

Several directions remain open. A first line of work concerns privacy-preserving multi-center learning: both TVTL-CLIP and CTVTL-CLIP naturally lend themselves to federated-learning settings, where only prompt parameters and lightweight support-network weights are updated and shared across institutions, limiting exposure of sensitive data while enabling cross-site generalization. A second direction is broadening applicability to additional medical imaging domains and label granularities, where the balance between visual prompting, textual

prompting, and conditional context may differ (e.g., radiology, histopathology, or ultrasound). Finally, real-time and resource-constrained deployment is a concrete target: prompt tuning already minimizes training cost, and further engineering can focus on inference optimization of the frozen backbone and the support network (e.g., quantization, distillation, and efficient batching) to enable practical use on commodity hospital hardware.

Overall, TVTL-CLIP establishes that jointly learning visual and textual prompts is sufficient to bridge a substantial portion of the gap between generic vision-language pretraining and domain-specific angiography classification. CTVTL-CLIP strengthens this result by demonstrating that instance-conditioned prompting can further improve robustness under realistic clinical variability, with the clearest gains emerging as task complexity increases, while preserving the same parameter-efficiency principles that make prompt-based adaptation attractive for medical settings.

7.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 240 hours of computation was performed on 3 NVIDIA A100 GPUs (thermal design power of 300 W), with total emissions estimated[1] to be 25.92 kg CO_2eq.

[1]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 8

FeDETR: a Federated Approach for Stenosis Detection in Coronary Angiography

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Amelia Sorrenti, and Federica Proietto Salanitri. Fedetr: A federated approach for stenosis detection in coronary angiography. In *International Conference on Image Analysis and Processing*, pages 189–200. Springer, 2023

8.1 Overview

Assessing the haemodynamic severity of coronary stenoses from coronary angiography is a clinically central step in the management of coronary artery disease, as stenosis-driven flow limitation is strongly associated with adverse outcomes and may drive the indication to revascularization. Guideline-recommended physiological assessment relies on invasive indices such as FFR and iFR [267, 183]. Although effective, pressure-wire based measurements introduce additional procedural time and cost, require dedicated equipment and operator expertise, and entail non-negligible invasiveness with potential complications. Moreover, clinical interpretation may be affected by inter-observer variability, especially in challenging cases where identifying the major stenosis is not straightforward [340]. These limitations have motivated substantial interest in automated, image-based estimation of FFR/iFR directly from routine angiography.

In the last decade, deep learning has enabled increasingly accurate assistance tools for cardiovascular imaging analysis and risk assessment [209, 261], including approaches

specifically designed for stenosis severity quantification in angiography [410, 409, 396]. Many of the most competitive solutions exploit multi-view acquisitions and, in several cases, depend on the concept of key frame selection, i.e., manually identifying frames where contrast opacification is maximal and vessel borders are best delineated. While these strategies can be effective, they typically require large, diverse datasets to generalize well and may still impose manual interaction on clinicians. In practice, the availability of sufficiently large angiography cohorts is limited by the distributed and privacy-sensitive nature of medical data: examinations are stored across institutions, governed by heterogeneous policies, and are rarely shareable at scale.

FL [231, 198, 202] provides a principled way to address this bottleneck by enabling collaborative training across multiple sites without centralizing patient data. In FL, each institution trains locally and shares only model updates, allowing the global model to benefit from the diversity of multi-center cohorts while keeping raw data within each hospital. This is particularly attractive for angiography-based stenosis assessment, where inter-site variability in acquisition systems, protocols, and patient populations may otherwise impair robustness, and where single-center datasets are often too small to support strong generalization.

In this context, we introduce **FeDETR**, the first federated detection transformer approach for stenosis severity assessment from angiography videos via estimation of FFR/iFR. In a centralized federation, each node trains a Detection Transformer (DETR) [46] on its local, heterogeneous dataset, while the central server aggregates knowledge by federating the backbone component of the network. This design aims to combine two complementary properties: (i) the strong representation power of transformer-based detection models for identifying clinically relevant regions and key frames, and (ii) the privacy-preserving, multi-site generalization benefits of FL. FeDETR targets the detection of haemodynamically significant stenosis in key frames and supports severity assessment through direct and indirect estimation of FFR/iFR-derived decision thresholds (FFR < 0.80 and iFR < 0.89) [267, 359, 75, 71].

We train and evaluate FeDETR on a multi-center cohort collected from five hospitals, comprising 1001 angiographic examinations. The proposed approach is assessed against state-of-the-art federated baselines, including FedAvg [231] and FedBN [202], to quantify the benefits of transformer-based detection under federated optimization and to validate the contribution of backbone federation in the presence of cross-site heterogeneity. Overall, this chapter positions FeDETR as a step toward scalable, privacy-preserving AI for coronary physiology assessment, aiming to reduce reliance on invasive pressure-wire measurements while improving robustness through multi-institutional learning.

8.2 Method

An overview of the proposed approach is shown in Fig. 8.1. In this scenario, a federation consists of a set of n peer nodes managed by a central server. Each node owns a private dataset consisting of *key frames* extracted from coronary angiography videos. This dataset is used to train a DETR model [46], as shown in Fig. 8.2.

During each round of centralized training, each node receives the backbone of the model from the server. The backbone is used to extract features from the images in the private dataset. After being flattened by the model and enhanced with position encoding, these features are fed into the encoder-decoder transformer. Each of the N output embeddings of the transformer decoder is then passed to a shared feed-forward network, which independently decodes it into bounding boxes and class labels, resulting in N final predictions.

At the end of each round, which involves multiple training iterations, the locally trained backbone from each device is sent back to the central server. The server performs the aggregation of the received backbones, and the entire process is repeated for the next round of training. This iterative process ensures collaboration between nodes while preserving data privacy and security.

8.2.1 Problem formulation

The proposed method operates in a federated setting, where K nodes jointly train for R rounds. At the end of each round, which consists of T local epochs, a communication with a central server is performed.

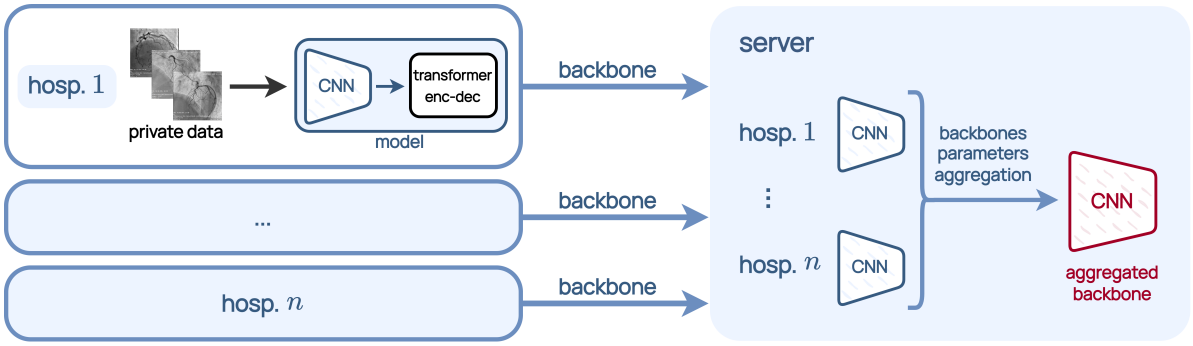


Figure 8.1: Each node trains a DETR model using its local dataset, composed of key frames extracted from coronary angiography videos. At the beginning of each round, the central server sends the aggregated backbone to each node, which uses it to extract features from the key frames. After each round, each node sends its locally trained backbone back to the central server, which aggregates them all.

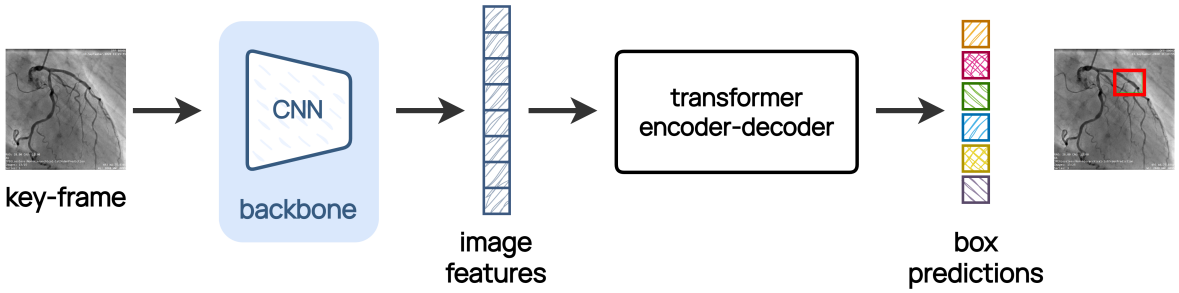


Figure 8.2: Given a key frame, the CNN backbone learns a 2D representation, which is then flattened by the model. Before passing these learned features to the transformer encoder, the model also extends them with positional encoding. The transformer decoder outputs some learned positional embeddings and feeds them into a feed-forward network that predicts classes and their corresponding bounding boxes.

Each node i trains a local model M_i on its own private dataset $\mathcal{D}_i = \{(\mathbf{x}_1, y_1, \mathbf{b}_1), (\mathbf{x}_2, y_2, \mathbf{b}_2), \ldots, (\mathbf{x}_n, y_n, \mathbf{b}_n)\}$, where $\mathbf{x}_j \in \mathcal{X}$ represents a sample image within the dataset, $y_j \in \mathcal{Y}$ represents the target object class (low-severity or mild/high-severity stenosis), and $\mathbf{b}_j \in \mathbb{R}^4$ represents bound box coordinates. Each local model $M_i = \langle g, \boldsymbol{\theta}_i \rangle$ is parameterized by $\boldsymbol{\theta}_i$, with $g : \mathcal{X} \to \mathcal{Y}$. Specifically, in the proposed approach, the K nodes train a local DETR model [46] whose architecture is illustrated in Fig. 8.2. Since the model consists of two parts, the model parameters $\boldsymbol{\theta}_i$ can be written as the union of the backbone and the encoder-decoder transformer parameters, $\boldsymbol{\phi}_i$ and $\boldsymbol{\omega}_i$ respectively.

Hence, at training time, each node i in the federation, locally optimizes a finite sum objective in the form:

$$\min_{\boldsymbol{\theta}_i} f(\boldsymbol{\theta}_i) \quad \text{with} \quad f(\boldsymbol{\theta}_i) \equiv f(\boldsymbol{\phi}_i, \boldsymbol{\omega}_i), \tag{8.1}$$

defining $f(\boldsymbol{\phi}_i, \boldsymbol{\omega}_i)$ as follows:

$$f(\boldsymbol{\phi}_i, \boldsymbol{\omega}_i) = \frac{1}{n} \sum_{j=1}^{n} f_j(\boldsymbol{\phi}_i, \boldsymbol{\omega}_i), \tag{8.2}$$

where $f_j(\boldsymbol{\phi}_i, \boldsymbol{\omega}_i) = \ell(g(\mathbf{x}_j), y_j, \mathbf{b}_j; \boldsymbol{\phi}_i, \boldsymbol{\omega}_i)$ denotes an object detection loss for $(\mathbf{x}_j, y_j)$ using the model parameters $\boldsymbol{\theta}_i$.

In contrast to current federated approaches, which are mostly based on the averaging of all model parameters, our method only federates the backbone part of the model. Indeed, at the end of each training round, the nodes in the federation apply the FedAvg algorithm [231] on the subset of parameters $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$, thus receiving an updated version of the backbone and resuming local training.

8.3 Experimental Evaluation

8.3.1 Dataset

The proposed method was trained and evaluated using a private dataset comprising of 1001 coronary angiographies from patients with CCS and ACS. The dataset was collected in five large hospitals between 2020 and 2022, and included 737 male and 264 female patients with an average age of 66.8 $\pm$ 9.26 years.

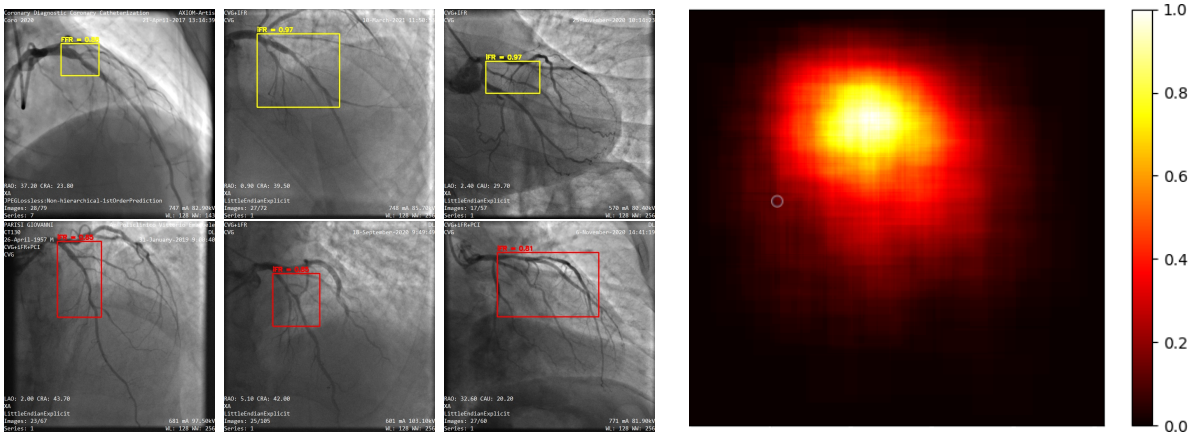
The coronary angiographies and the physiological measurements were carried out in accordance with the standard clinical practice. Then, two experienced cardiologists annotated the key frames in the angiography data. Each angiography was evaluated using invasive physiological assessment with iFR, FFR, or both. Specifically, FFR values were available for 613 patients (61.2%), iFR values for 667 patients (66.6%), and a subset of 279 patients (27.8%) had both FFR and iFR data. Cardiologists identified the major stenosis in each exam and labeled it as hemodynamically significant if the FFR was less than 0.80 [267, 359, 75] or if the iFR was less than 0.89 [267, 71]. Consequently, the 40.5% of patients has been labeled as positive, resulting in a significant class imbalance.

8.3.2 Training procedure

Each node trains a DETR model [46], consisting of a ResNet-50 [141] backbone and an encoder-decoder transformer. The model was pre-trained on the COCO 2017 dataset [207]. The DETR models were trained by minimizing the Hungarian loss [46] using the AdamW [218] optimizer with initial learning rates set to 10^{-4} and 10^{-5} for the transformer and the backbone, respectively, and weight decay set to 10^{-4}. The learning rate is then updated using a scheduler with the multiplicative factor γ set to 0.1. Gradient clipping was also applied with a maximum gradient norm of 0.1.

Both the encoder and the decoder consist of six layers characterized by a standard architecture. Their self-attention and cross-attention modules have eight heads, each receiving input embeddings of size 256. In addition, the feed-forward network (FFN) of these modules has a dimensionality of 2048. An additive dropout of 0.1 is applied after each multi-head attention and FFN and before performing the layer normalization. The output of the encoder is supplemented with a sinusoidal positional encoding before being passed to the decoder. The number of object queries of the transformer decoder is 20.

Data augmentation is performed by consecutively and randomly applying horizontal and vertical flips, and 90-degree rotations. Scale augmentation is then applied to the images, resizing them so that the shortest side is 512 pixels and the longest side is 1333 pixels or less.



(a) (b)

Figure 8.3: Figure 8.3a shows representative examples of *low severity stenoses* (illustrated in the top row) and *mild or high severity stenoses* (depicted in the bottom row). Figure 8.3b presents an averaged heatmap, which is derived from the collective bounding box positions across all samples.

Furthermore, zero padding is also added to the images, making them rectangular in order to shift the detection area. Since bounding boxes are usually positioned in the upper left corner or centrally, as demonstrated by the heatmap in Fig. 8.3b, the additional padding should prevent the model from learning a location bias.

8.3.3 Results

To evaluate FeDETR, we compare it against baseline federated approaches, namely FedAvg [231] and FedBN [202], using standard object detection metrics, including ACC, PPV, TPR, and IoU. Results are presented in Tab. 8.1.

The federated training approach consistently achieves significant accuracy gains over baseline DETR models that utilize a combined dataset from all nodes (*joint*), and even more compared to local training alone (*standalone*). This outcome underscores that training on smaller individual node datasets outperforms a unified dataset in a single node. Furthermore, when comparing federated training on individual node datasets to training on a single node's dataset, there is a notable 12-point increase in accuracy. This suggests that federated training on node-specific data surpasses training on a single consolidated dataset, possibly due to large data distribution shifts between nodes.

The comparison between the three federated approaches, i.e. FeDETR, FedAvg [231] and FedBN [202], shows that all of them achieve similar performance in both classification accuracy and object detection metrics. In particular, FeDETR shows generally improved performance at a lower number of rounds, highlighting better convergence properties. Also,

Table 8.1: Performance comparison between the proposed approach, FedAvg and FedBN, for different combinations of number of rounds and epochs, on the target metrics. Standard deviations refer to inter-node variability.

Rounds	**Epochs**	**Methods**	**PPV**		**TPR**		**IoU**		**ACC**
			low	*high*	*low*	*high*	*low*	*high*	
-	200	*joint training*	69.8	61.9	80.2	48.1	52.2	50.8	67.3
-	200	*standalone*	64.8	58.1	68.7	43.2	43.9±2.3	42.7±3.8	58.5±3.2
25	5	FedAvg [231]	55.4	29.5	78.0	25.9	53.1±4.1	42.0±11.1	72.2±4.4
		FedBN [202]	57.5	27.9	73.3	34.1	51.8±5.2	54.3±1.4	71.8±4.1
		Ours	57.9	33.5	73.4	33.9	51.2±1.8	51.3±6.8	70.9±3.5
	20	FedAvg [231]	55.2	28.1	77.3	25.9	51.0±2.6	40.8±11.2	71.8±4.6
		FedBN [202]	59.1	47.1	80.5	25.3	54.5±3.2	52.1±3.8	69.0±1.9
		Ours	62.0	47.9	77.8	31.2	53.3±4.0	37.3±9.4	70.2±3.0
50	10	FedAvg [231]	57.8	31.7	77.3	32.9	51.4±2.9	44.6±8.5	74.3±4.5
		FedBN [202]	58.3	29.3	74.1	34.4	55.7±3.9	52.1±3.6	72.4±4.0
		Ours	59.8	38.7	68.1	41.1	51.1±3.4	40.2±5.2	71.0±3.4
	5	FedAvg [231]	58.3	28.9	74.0	35.3	53.2±4.3	54.9±2.9	73.1±4.3
		FedBN [202]	63.0	73.3	85.2	32.8	52.8±3.4	54.0±1.3	71.1±2.3
		Ours	64.4	66.3	78.9	39.8	46.2±4.2	44.2±5.2	69.4±1.6
	10	FedAvg [231]	62.0	56.1	83.4	28.3	55.1±4.3	47.5±10.2	69.2±2.0
		FedBN [202]	63.5	72.2	94.9	30.4	55.5±3.7	54.3±4.4	69.5±3.2
		Ours	66.9	66.8	81.1	46.5	51.2±2.9	46.9±6.9	73.6±3.8
	20	FedAvg [231]	59.3	49.8	74.1	37.3	56.6±3.1	52.8±2.1	73.9±4.1
		FedBN [202]	61.2	47.6	85.4	27.8	58.5±5.4	48.3±4.8	66.0±1.5
		Ours	67.6	49.8	69.5	47.9	50.3±3.5	50.0±2.9	70.4±3.7

while FeDETR seems to achieve better scores in precision rather than recall, we can observe a general trend of improvement on mild/high-severity stenosis, which can be a critical factor for diagnostic purposes. On the other hand, FeDETR shows worse performance on bounding box prediction, based on the IoU metric.

In particular, FeDETR employs a model that shares only 15% of the total model's parameters, specifically the backbone, omitting the transformer component. This strategic parameter sharing maintains privacy concerns (since weight sharing can be subject to malicious attacks aimed at reconstructing model inputs [111, 424]) while still delivering competitive performance.

These results collectively underscore the effectiveness of federated training in enhancing accuracy and performance across various scenarios, showcasing its potential for collaborative and privacy-preserving machine learning.

8.4 Discussion

This work investigates a federated learning strategy for coronary stenosis assessment that combines the privacy-preserving advantages of multi-center optimization with the representation power of DETR models [46]. By explicitly incorporating invasive physiology targets, FFR and iFR, the proposed framework is designed to better align visual cues in coronary angiography with haemodynamic severity, rather than relying only on morphology-driven surrogates. In practical terms, this shifts the objective from purely anatomical interpretation toward a physiology-informed assessment, which is closer to the decision process used to guide revascularization.

Methodologically, FeDETR adopts a centralized federation where each site trains a DETR model to identify high-severity stenoses on key frames extracted from angiography videos, while the server aggregates only the backbone parameters at the end of each communication round. This choice reflects a balance between global generalization and local specialization: federating the backbone encourages a shared feature extractor that can absorb cross-site variability, whereas keeping non-backbone components local helps preserve site-specific adaptation when the underlying data distributions are non-i.i.d. Such a design is particularly relevant in angiography, where institutional differences in acquisition devices, protocols, contrast dynamics, and patient populations can induce systematic shifts that degrade centralized models trained on limited single-center cohorts.

From a clinical translation perspective, the federated setup directly addresses a core barrier to scaling learning-based stenosis assessment: the difficulty of collecting and sharing sufficiently large datasets across hospitals under realistic governance constraints. By ensuring that patient data remain on-premise while still enabling collaborative training, the proposed approach provides a pathway to exploit multi-center diversity without sacrificing privacy or data ownership. In this sense, FeDETR is not only a modeling contribution but also an operational one, as it aligns with how hospitals can realistically participate in joint development of decision-support systems.

Empirically, evaluation on a real multi-center dataset collected from five hospitals indicates that the federated DETR approach is competitive with established federated baselines and, in particular, supports robust discrimination between mild and high-severity stenoses. Importantly, these results are obtained under realistic heterogeneity, which strengthens the plausibility of deployment in clinical scenarios where domain shift is the norm rather than the exception. The overall evidence suggests that combining transformer-based detection with backbone federation is a viable approach to preserve performance while improving robustness across sites.

Beyond the specific application to coronary stenosis, this study also serves as an early example of federated detection transformers for computer vision tasks where local datasets are inherently non-i.i.d. The broader implication is that partial federation strategies, here implemented through backbone aggregation, can offer a practical mechanism to share transferable representations while retaining local expertise in task-specific or site-specific components, especially when strict homogenization of data is impossible.

Several limitations and extensions remain open. First, the current pipeline assumes the availability of key frames; automating key-frame selection directly within the learning loop would reduce reliance on pre-processing and further improve usability in routine workflows. Second, extending the framework to additional contrast-enhanced examinations (e.g., other angiographic procedures or imaging protocols where contrast dynamics is informative) could test the generality of the approach and its ability to learn transferable haemodynamic priors under federated optimization. Finally, future investigations should focus on external prospective validation, robustness under distribution shifts not represented in the training federation, and operational aspects of real deployments (communication constraints, federation governance, and monitoring of site-wise performance).

8.5 CO_2 Emissions Related to Experiments

Experiments were carried out using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A total of 300 hours of computation was performed on multiple GPUs (i.e., two NVIDIA RTX 3090, two NVIDIA Ampere A100, and two NVIDIA RTX A6000 GPUs), with total emissions estimated[1] to be 38.88 kg CO_2eq.

[1]Estimations were performed using the https://mlco2.github.io/impact#compute [189].

Chapter 9

A Cardiac Biometric P.P.E. as Emergency Stop Device

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Paolo Soda, Simone Palazzo, and Concetto Spampinato. Biometric safety system and method for estimating a user hazard condition and triggering an emergency stop, 2027. IT Patent

9.1 Evaluation, Calibration, and Clinical Considerations

9.2 System Architecture

9.3 Signal Acquisition, Preprocessing, and Signal-Quality Control

9.4 Feature Extraction and Calibration

9.5 Dual-Channel Anomaly Detection and Decision Fusion

9.6 Safety Chain, Actuation Logic, and Industrial Deployment

9.7 Privacy, Logging, and Auditability

9.8 Validation Targets, TRL Roadmap, and Standards Alignment

9.9 Discussion, Limitations, and IP Disclosure Boundaries

Part II

AI-Driven Cognitive and Behavioral Monitoring

Chapter 10

Overview of Automated Cognitive Assessment

Cognitive decline and neurodegenerative disorders represent a major and growing societal challenge, with significant impact on patients, caregivers, and healthcare systems. Early detection and longitudinal monitoring are essential to support timely interventions, personalize care pathways, and evaluate treatment effects. In practice, however, frequent cognitive assessment is difficult to scale. Standard screening and neuropsychological evaluations require trained personnel, dedicated clinical time, and controlled environments, and they can be influenced by subjective scoring, patient fatigue, and variability across assessors. These constraints create a gap between clinical need and operational feasibility: many individuals are assessed late in the disease trajectory or monitored infrequently, limiting opportunities for early support and continuous care. Automated and semi-automated approaches can help reduce this gap by enabling more frequent, objective, and scalable assessment, including outside specialized clinical settings.

Within the framing of this dissertation, cognitive monitoring also reflects a broader deployment reality: real-world impact is often limited by the feasibility of acquiring suitable inputs, rather than by the availability of powerful models. If clinically informative signals can be collected through non-invasive and widely available modalities, automated assessment becomes more accessible and deployable, supporting screening and follow-up at scale.

Clinical instruments such as MMSE [105] and MoCA [265] provide structured evaluations across cognitive domains including memory, attention, language, executive function, and visuospatial abilities. While these instruments are clinically useful, they are typically administered episodically and rely on human interaction and scoring. Automated cognitive assessment aims to estimate clinically relevant outcomes (e.g., screening scores or impairment risk) from measurable signals that correlate with cognitive function. These signals may

include behavioral cues such as facial expressions, micro-movements, head pose, eye-gaze patterns, and reaction times; speech and language markers such as prosody, lexical richness, hesitations, and semantic coherence; motor patterns reflecting fine motor behavior and coordination during task execution; and physiological signals, including heart rate variability or sleep-related markers, often used as complementary indicators in longitudinal monitoring.

This dissertation focuses on scalable behavioral signals, with emphasis on vision-derived cues and complementary wearable time series, motivated by their non-invasive nature and the possibility of acquisition using commodity devices such as webcams, tablets, eye-tracking-capable systems, and consumer smartwatches. The central question is whether subtle, everyday signatures of behavior can provide sufficient information for meaningful cognitive assessment and monitoring, and how learning pipelines can be designed to remain reliable under realistic acquisition variability, missingness, and inter-subject heterogeneity.

Automated cognitive assessment introduces methodological challenges that differ from many standard vision tasks. First, cognitive-related behavioral cues can be subtle, temporally sparse, and strongly confounded by inter-subject variability. Micro-movements may be difficult to observe without enhancement, and gaze trajectories depend on both task design and user interaction. Second, labeled data are often limited and heterogeneous: collecting cognitive assessment datasets requires clinical protocols, and data can vary by site, population, and acquisition device, increasing overfitting risk and strengthening the need for careful validation and regularization. Third, temporal structure and context are fundamental: many cues are meaningful only over time, and models must handle dynamic patterns, irregular sampling (common in gaze data), and variable session length. Fourth, clinical validity and evaluation must align with decision pathways: a regressor estimating a MoCA score should also be evaluated against clinically meaningful cut-offs, and outputs should be assessed for calibration and robustness under distribution shift. Fifth, ethical and privacy considerations are central to deployment, because monitoring pipelines may involve identifiable data such as faces and backgrounds; privacy-preserving collection, careful governance, and, where feasible, modality choices that reduce identifiability are therefore relevant design constraints rather than optional add-ons.

A further limitation concerns the scale and origin of the datasets used in this part. The studies on facial micro-dynamics and eye-gaze modeling rely on limited, project-specific cohorts acquired under controlled or semi-controlled protocols. These datasets are appropriate for methodological validation and for assessing whether subtle behavioral signals can support cognitive score estimation, but they are not sufficient to fully characterize generalization across clinical sites, devices, populations, and acquisition protocols. Similarly, the wearable-sensing study provides evidence under realistic daily-life missingness and heterogeneity, but remains

tied to the data distribution and labeling structure of the challenge dataset. Therefore, the results in Part II should be interpreted primarily as proof-of-concept evidence for scalable multimodal cognitive monitoring. Full clinical generalization will require larger multi-site cohorts, external validation, population-aware evaluation, and standardized acquisition protocols that enable reproducible comparison across devices and settings.

Part II develops three complementary contributions that reflect these constraints and opportunities. The first investigates methods for magnifying facial micro-movements to enhance subtle behavioral signals and integrate them with task-related descriptors for cognitive evaluation. The second explores transformer-based modeling of eye-gaze traces to estimate MoCA scores from structured tasks, emphasizing temporality, irregular sampling, and evaluation protocols aligned with remote assessment. The third analyzes long-term wearable sensing in unconstrained daily-life conditions, leveraging multimodal smartwatch time series to address both subject identification and personalized relapse detection in the psychotic spectrum under realistic missingness and acquisition heterogeneity. Together, these studies support the broader thesis message that clinically useful progress in cognitive monitoring requires coupling strong representation learning with multimodal sensing pipelines that remain practical, accessible, and compatible with real-world constraints.

Chapter 11

Related Work in Cognitive Monitoring

Automated cognitive assessment increasingly leverages behavioral and physiological signals as non-invasive proxies of neurocognitive status, with the aim of enabling scalable screening and longitudinal monitoring outside specialized clinical settings. Prior research spans vision-based facial analysis, eye-gaze and eye-tracking, and long-term wearable sensing, alongside a shift from handcrafted descriptors to deep learning and attention-based architectures. This chapter reviews these lines of work and outlines methodological gaps that motivate the contributions of Part II.

Standardized screening tools such as MMSE and MoCA provide widely adopted quantitative measures of cognitive function across multiple domains, including attention, memory, visuospatial abilities, and executive functioning [105, 265]. Despite their clinical utility, traditional administration remains time-consuming, dependent on trained personnel, and sensitive to examiner variability and testing context. These limitations have driven interest in automated estimation of cognitive scores and/or impairment categories using digital biomarkers captured during structured tasks, serious games, or passive observation, improving scalability and repeatability while reducing burden on clinics and patients.

A major research direction investigates facial behavior as a proxy of cognitive state. Facial activity is a rich communication channel and has been studied through structured computational representations such as the Facial Action Coding System (FACS) [89]. Prior works have connected facial patterns to attentional processes [128, 208], affective dynamics [386], pain expression [187], and language-related pathologies [232], supporting the broader hypothesis that facial behavior correlates with neurological and cognitive conditions. In neurocognitive decline, studies have examined facial expression dynamics in mild cognitive impairment (MCI) and Alzheimer's disease (AD), including analyses that exploit 3D sensors such as Kinect [114]. While macro-expression dynamics can be informative, several works argue that spontaneous and involuntary components of facial activity can provide additional

discriminative information, especially when the observable signal is subtle and inter-subject variability is strong.

Within facial analysis, a specific line of work focuses on micro-expressions, subtle and brief involuntary movements that may occur when subjective states are suppressed or when neurocognitive disorders affect emotion regulation and control [90, 21, 164]. Micro-expressions typically last from tens to a few hundred milliseconds [398], which makes them difficult to capture reliably with standard cameras and manual observation. This has motivated algorithmic strategies that amplify subtle facial motion in conventional videos. In particular, Eulerian and phase-based video magnification methods can reveal small motion components that are otherwise hard to observe [392]. In cognitive monitoring pipelines, such motion enhancement can be paired with downstream learning models to extract micro-dynamics descriptors from low-cost recordings, reducing reliance on specialized acquisition and enabling more scalable deployment.

Eye-gaze and eye-tracking constitute another prominent non-invasive modality for probing cognitive function, especially in aging populations and subjects at risk of MCI and AD [22]. Gaze-derived features such as fixation duration, saccade amplitude, scanpath regularity, and pupil dynamics have been associated with attention, memory, and executive functioning. Lagun et al. reported that gaze features can discriminate MCI from healthy controls [190], and clinical reviews have highlighted relationships between oculomotor abnormalities and neurodegenerative processes [8]. Beyond passive observation, several studies embed eye-tracking into structured cognitive tasks. For example, eye-tracking variants of the Trail Making Test have been explored to predict cognitive impairment and report associations between gaze behavior and clinical outcomes [142]. Multimodal settings that combine gaze with speech have also been proposed for remote MCI classification [160]. More broadly, supervised [190] and unsupervised learning [304] on gaze patterns reinforce the feasibility of gaze-based screening, while raising open questions on task dependence, robustness to sampling irregularities, and transferability across devices and cohorts.

Long-term wearable and smartphone sensing has also emerged as a practical route to quantify behavioral and physiological patterns in everyday life, often framed within digital phenotyping and passive monitoring for psychiatry [223, 274]. In the psychotic spectrum, prospective studies have investigated relapse prediction using passively collected smartphone signals and longitudinal behavioral traces, showing that relapse-related deviations can be captured under unconstrained acquisition regimes [62]. Systematic reviews further summarize the state of passive sensing for relapse-related outcomes in schizophrenia and emphasize persistent limitations such as cohort scale, confounding, and heterogeneity in acquisition and labeling protocols [330]. Complementary directions investigate wearable-only or wearable-first

streams and deep representation learning, including autoencoder-based pipelines, to model subject-specific dynamics and detect clinically meaningful deviations over time [397]. These long-horizon streams introduce challenges that are less pronounced in short, laboratory-style recordings, including heterogeneous sampling, missing segments, and strong inter-subject variability. This has motivated missingness-aware sequence modeling and imputation-oriented architectures for multivariate time series [52, 44], as well as attention-based temporal encoders designed for long contexts and multi-sensor integration [205, 420].

The modeling landscape for behavioral and physiological signals has progressively shifted from handcrafted descriptors to end-to-end learning. Early systems for cognitive score prediction often relied on manually engineered features paired with classical predictors. For MoCA estimation, examples include Random Forest models built on social and biomarker variables [206] and pipelines based on handcrafted interaction features such as keystroke dynamics [146]. While interpretable, such pipelines frequently demand extensive feature design and may generalize poorly across heterogeneous tasks, acquisition setups, or user populations.

Deep learning enabled more direct modeling of sequential behavioral data. Recurrent architectures such as RNNs, LSTMs, and GRUs model temporal dependencies by iteratively updating a latent state [24, 145, 58] and have been applied to gaze sequences and related time series, including personalized frameworks that infer latent states from low-level gaze features [173]. However, many efforts target transient affective states rather than structured cognitive scoring, and they often underutilize spatial grounding or task semantics. Alternative designs process gaze as image-like representations (for example, heatmaps) through CNNs and then model temporal evolution using gated recurrent units or related modules. VTNet, for instance, uses dual branches for scanpaths and temporal gaze streams and reports encouraging results in early AD detection [345], but it depends on task-specific fusion and separate processing pipelines that can complicate deployment and robustness analysis.

More recently, transformers introduced self-attention as an effective mechanism for modeling long-range dependencies in sequential data [373], with adaptations to time-series, vision, and biomedical signals. In computer vision, ViT demonstrated that tokenized representations can learn strong visual features and support flexible sequence modeling [85]. Transformer-style modeling has also been explored for gaze-related applications, including attention-based frameworks for gaze-driven interaction [255]. These trends are relevant for automated cognitive assessment because many behavioral signatures are inherently spatiotemporal. In such settings, what matters is not only a temporal pattern but also its spatial organization relative to task-relevant regions, and attention-based encoders provide a natural mechanism to integrate these dimensions while retaining flexible handling of irregular sampling and

variable sequence length.

Across facial micro-dynamics, eye-gaze modeling, and wearable sensing, recurring methodological gaps motivate the contributions of Part II. First, robust extraction of subtle behavioral signals under realistic acquisition constraints remains difficult, especially with low-cost sensors, motion artifacts, and heterogeneous environments. Second, there is a need for end-to-end spatiotemporal modeling that preserves context and supports structured score estimation (MMSE, MoCA) rather than only categorical impairment detection. Third, long-term monitoring requires pipelines that remain stable under missingness, heterogeneous sampling, and strong inter-subject variability, often calling for personalization and explicit missingness-aware modeling. The contributions in Part II address these gaps through (i) a pipeline that enhances facial micro-movements from standard videos and integrates affective and task-related descriptors for MMSE estimation, (ii) transformer-based modeling of eye-gaze traces for MoCA estimation in structured testing settings, and (iii) a long-term wearable sensing study that evaluates deep time-series pipelines under realistic missingness and heterogeneity, covering both subject identification and personalized relapse detection.

Chapter 12

Magnifying Facial Micro-movements for Cognitive Evaluation

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Federica Proietto Salanitri, Lisa Passarello, Alberto Sardella, Laura Messina, Manuela Pennisi, Daniela Giordano, Simone Palazzo, and Concetto Spampinato. Magnifying facial micro-movements for cognitive evaluation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024

12.1 Overview

Objective cognitive assessment is essential for early detection and management of cognitive decline, yet widely adopted instruments such as MMSE [105] still depend on face-to-face administration and manual clinician scoring. This limits scalability, repeatability, and longitudinal follow-up, motivating automated alternatives that can be deployed in low-burden or remote settings. In this work, we investigate facial behavior recorded during structured cognitive tasks and serious games [84] performed on a tablet as a non-invasive behavioral signal for estimating MMSE outcomes. A practical challenge is that potentially informative facial cues are often subtle and short-lived, and may be attenuated by commodity acquisition conditions, including limited temporal resolution, sensor noise, head-pose variability, and illumination changes.

To address this limitation, we introduce a multimodal deep learning pipeline that explicitly enhances and exploits facial micro-dynamics. Specifically, we apply Phase-Based Video Magnification (PBVM) [392] as a visual preprocessing step to amplify fine-grained facial motion components from standard videos, enabling more reliable extraction of micro-movement

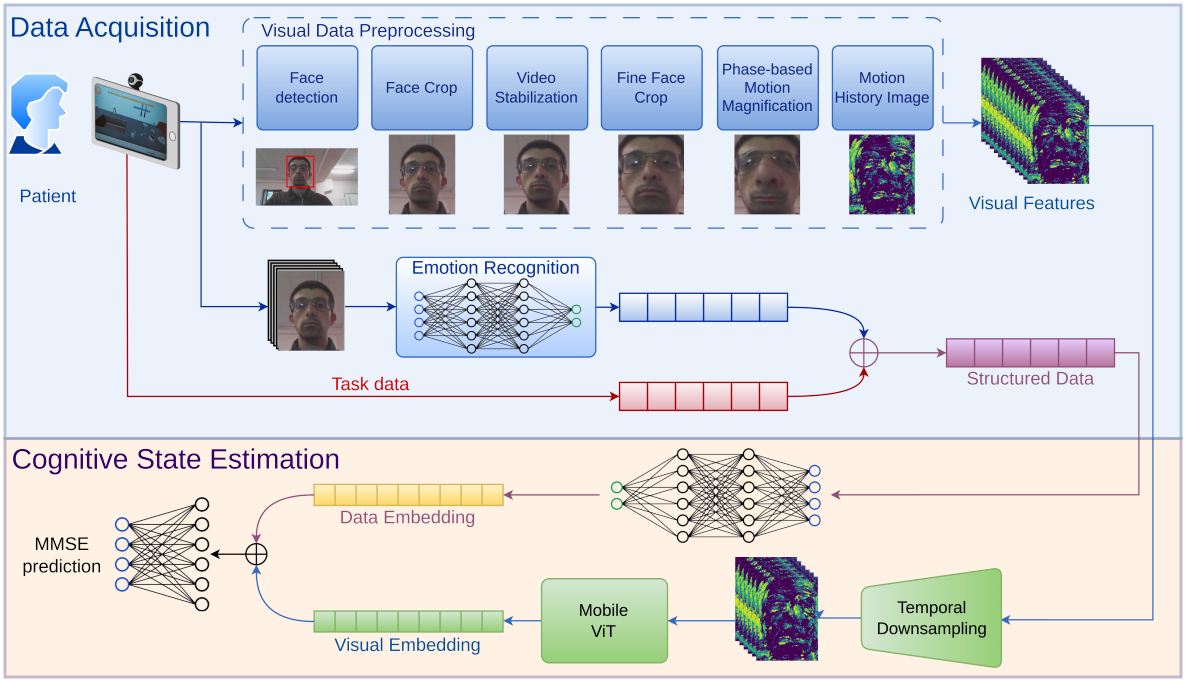


Figure 12.1: Overview of the proposed approach. Facial micro-movement features are extracted through a visual pre-processing pipeline and combined with emotional states and performance metrics on cognitive tasks to predict a subject's MMSE score

descriptors without requiring specialized cameras. The resulting features are then integrated with complementary contextual signals collected during the session, including (i) emotional state estimates inferred from the video stream and (ii) performance indicators derived from task interaction. The model leverages this fused representation to predict MMSE scores (or score-consistent categories), providing evidence of a measurable association between facial micro-movements and cognitive status. Overall, the study supports the feasibility of automated video-based analysis, coupled with motion magnification and task context, as a scalable proxy for objective cognitive assessment.

12.2 Method

12.2.1 Architecture Overview

The architecture of the proposed approach consists of two main stages, as illustrated in Fig. 12.1: a *visual pre-processing* pipeline, which analyzes facial features from a video to capture and emphasize facial micro-movements, and a *cognitive state estimation* deep network, which combines facial micro-expressions with emotional states and performance metrics on cognitive tasks to predict the subject's MMSE score.

12.2.2 Visual Pre-processing

The pre-processing of video data is a critical step in our study, as it ensures that the input to our deep learning model is suitable for the analysis of facial micro-movements and expressions. This stage is particularly important because motion magnification can amplify not only subtle facial dynamics, but also nuisance factors such as macro-movements, illumination changes, tracking errors, and sensor noise. Our pre-processing pipeline therefore aims to reduce global variability before magnification and to constrain the analysis to facial regions that are more likely to contain task-relevant micro-dynamics. The pipeline receives a video clip of a subject performing a cognitive task and consists of several stages:

1. **Face detection**: We apply a Multi-task Cascaded Convolutional Network (MTCNN) [412] to detect faces within the video frames. MTCNN is a three-stage deep learning model that effectively identifies facial structures with high accuracy, even under different lighting and pose conditions.

2. **Main face selection**: If multiple subjects are detected, we use a heuristic based on proximity to the tablet and the size of the detected face to select the primary subject, presumed to be the patient.

3. **Emotion recognition**: We adopt HyperExtended LightFace [326] to perform frame-by-frame emotion recognition, classifying each detected face into one of seven emotional states (anger, disgust, fear, happiness, sadness, surprise, neutral). This stage produces a 7D vector, where each component encodes the prevalence of each emotion as the ratio of video frames in which that emotion is recognized.

4. **Face crop**: In order to maintain a consistent focus on the face region throughout the analysis and remove unnecessary information, we crop the subject's face in each frame. The crop region extends the face bounding box by 25% vertically and horizontally, to allow for slight movements.

5. **Video stabilization**: We employ the Good Features to Track (GFTT) detector [331] from the VidStab library to mitigate the effects of involuntary patient motion and stabilize the background. This technique corrects for frame-to-frame jitter and provides a more stable foundation for accurate motion magnification. By reducing global displacement before magnification, this step helps limit the amplification of macro-movements and camera-induced oscillations.

6. **Fine face crop**: Post-stabilization, we perform a second round of face detection using MTCNN to refine the region of interest, cropping the face bounding box. This step

ensures that the focus remains tightly on the facial area after any adjustments made during stabilization.

7. **Motion magnification**: We apply a Phase-Based Motion Magnification algorithm [379] to amplify subtle facial movements that are otherwise imperceptible. Compared with a direct pixel-level amplification strategy, phase-based magnification is better suited to subtle motion analysis because it operates on local phase variations in a multi-scale representation and can be combined with temporal filtering to emphasize motion components within a selected frequency band. This is useful in the present setting because very slow global drifts, abrupt whole-scene changes, and high-frequency sensor noise are less likely to be amplified when they fall outside the selected band-pass range.

   The algorithm decomposes video frames into a complex steerable pyramid and computes the local phase over time at every spatial scale and orientation. These phases are temporally bandpassed to isolate specific temporal frequencies, corresponding to motion in different spatial scales and orientations. Magnified motion is then computed by amplifying the bandpassed phases by a factor α. The entire procedure can be summarized as follows:

$$\Phi_{\text{amplified}} = \Phi + \alpha \cdot (\Phi - \Phi_{\text{filtered}}), \tag{12.1}$$

   where Φ is the original phase of the image pyramid coefficients, Φ_{filtered} is the temporally filtered phase, and α is the amplification factor. The amplified phase $\Phi_{\text{amplified}}$ is then used with the original amplitude to reconstruct the motion-magnified video signal.

   Although this stage mitigates part of the variability introduced by macro-movements and camera instability, it does not remove all possible artifacts. Large head rotations, abrupt illumination changes, and noise components that fall within the amplified frequency band can still be enhanced. For this reason, the method is expected to be more reliable under controlled or semi-controlled acquisition conditions, where face visibility, illumination, and camera stability are sufficiently consistent.

8. **Motion History Image**: To capture the temporal aspect of facial movements, we generate a Motion History Image (MHI) [31] for each 3-second window of the video. MHI is a visual representation of motion where more recent movements are brighter, and older movements fade over time: it is updated at each time step t by overlaying the binary motion image $D(x, y, t)$ onto the MHI from the previous time step $H_{t-1}(x, y)$.

The updating rule for the MHI is as follows:

$$H_t(x,y) = \begin{cases} \tau & \text{if } D(x,y,t) = 1, \\ \max(0, H_{t-1}(x,y) - \delta) & \text{otherwise.} \end{cases} \tag{12.2}$$

where τ is a constant intensity value assigned at (x, y) locations where motion is detected (i.e., where $D(x, y, t) = 1$) and δ is the decay rate, which controls how quickly the intensity value decreases over time in the absence of new motion. Using MHI allows for efficiently aggregating motion information into a compact representation, increasing the temporal context for MMSE prediction. Examples of MHI outputs are reported in Fig. 12.2.

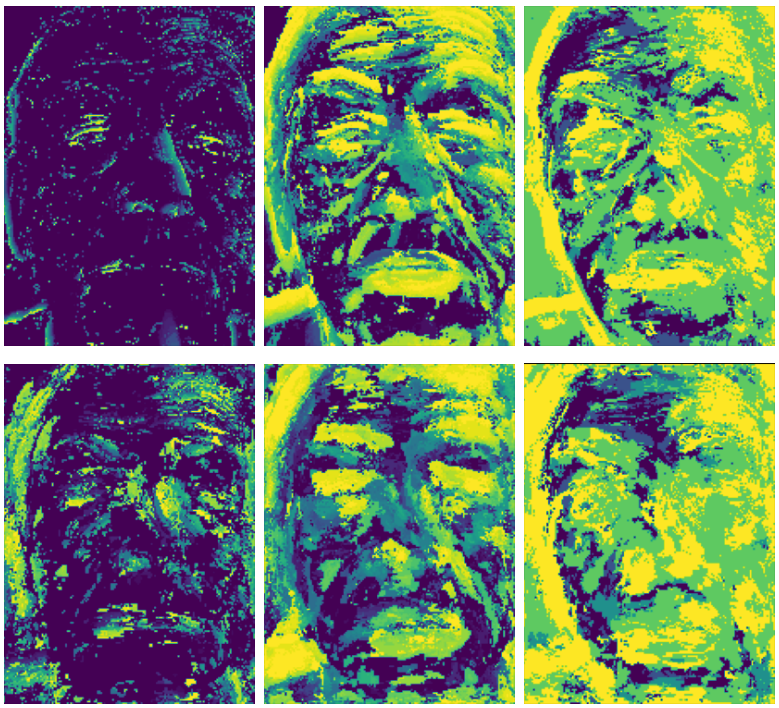


Figure 12.2: Examples of motion history images at different temporal granularities. The three columns represent the effects using windows of 1, 10 and 60 seconds. The top row shows the videos after face stabilization and cropping, and the bottom row shows the same videos after motion magnification

12.2.3 Cognitive State Estimation

In order to effectively exploit the rich data collected while a subject carries out cognitive tasks on a tablet device, the proposed cognitive state estimation model features a two-branch

processing architecture. The first branch is designed to manage the structured data that comes directly from the interactive tasks, including standardized indices for sessions and runs (indicating the degree of familiarity of the subject with the exercises and the number of exercises performed within the session, reflecting tiredness), performance scores from the cognitive exercises, an identifier for each specific exercise, and the duration of the exercises. This information is also integrated by emotion prevalence vectors, described above. We concatenate these inputs into a 12-dimensional vector and feed it to a fully-connected network, generating a 256-dimensional embedding that encapsulates structured task data.

The visual branch processes variable-length video frames through a trainable *temporal downsampling module*, which receives the sequence of motion history images and feeds them to a cascade of two 3D convolutions operating on the temporal dimension alone. As output this module provides an image-like tensor of resolution $H \times W \times T$, with H, W and T made independent of the original input size by means of adaptive max pooling. These features are then fed to a pre-trained Mobile ViT [203]; the class token is used as an aggregate representation and projected to a 256-dimensional visual embedding.

Both 256-dimensional embeddings from structured data and visual features are concatenated to form a 512-dimensional feature vector, which is then processed through two fully-connected layers to predict the subject's MMSE score.

12.3 Experimental Results

12.3.1 Dataset

The dataset was collected from a clinical trial involving 23 patients undergoing rehabilitation sessions. The study sample consisted of 13 male and 10 female participants from two different hospitals, providing a variety of clinical representations.

The number of sessions per patient ranged from 5 to 13, each lasting approximately 40 minutes and taking place on different days. This variability reflects individual differences in cognitive decline and rehabilitation progress. Each rehabilitation session consisted of 8 to 17 exercises, indicating the patients' level of participation in the rehabilitation program. The frequency and duration of the sessions were adjusted based on the patients' cognitive abilities and speed. Exercises were selected from a set of seven distinct cognitive tasks, categorized into four serious games related to activities of daily living and three additional cognitive tasks, providing a variety of stimuli and cognitive challenges that are critical for evaluating the rehabilitation progress of patients.

Data collection involved recording videos of the patients during each session. This provided

a dual modality for a comprehensive analysis of patient interaction and cognitive performance. A total of 162 sessions was recorded, resulting in 2294 individual exercise runs, each lasting between 1 and 10 minutes.

The patient sample showed mild cognitive impairment, with MMSE scores ranging from 20 to 28. We also recorded the score obtained in each question of the MMSE described in [105]. The dataset's range of MMSE scores enables a nuanced analysis of cognitive responses in correlation with varying degrees of cognitive function.

12.3.2 Training and Evaluation

We split the dataset into training, validation and test sets with proportion 56/22/22, such that all sessions of the same subject belong to a single split.

We train the proposed model using the AdamW [218] optimizer with a weight decay of $5 \cdot 10^{-4}$, a learning rate of $5 \cdot 10^{-5}$, and a batch size of 12. Parameters α, τ and δ are set, respectively, to 600, 255, 5. The temporal downsampling module provides an output of 16 frames at 224×224 resolution. Following the literature on the prediction of non-categorical ordered values [371], we minimize a weighted cross-entropy loss, accounting for the class imbalance within the dataset. Class labels correspond to MMSE scores of patients in our dataset (20 to 28). Training was conducted until convergence, and model selection is performed based on validation accuracy.

Results on MMSE estimation are reported in Tab. 12.1, in terms of balanced accuracy, sensitivity, specificity and F1 score. Performance values are reported as mean $\pm$ standard deviation over repeated runs under the adopted evaluation protocol, in order to provide an estimate of variability and experimental stability. Employing task data and emotions alone already provides above-chance prediction accuracy: the former encodes information on the capability of a subject to solve a task, which affects the duration of the exercise; the latter reflects the serenity or frustration of the subject, which correlates with the relative task complexity. Visual features alone, however, outperform the usage of task data and emotions, demonstrating the effectiveness of facial expressions in assessing cognitive capabilities. Importantly, the model's accuracy benefits from motion magnification, which adds an improvement of 3.5 points in accuracy and 7.5 points in F1 score with respect to the combination of both visual and structured features, confirming the contribution of facial micro-movements and their relevance to MMSE estimation under the considered protocol.

Besides evaluating our method on MMSE estimation, we extend our analysis to the prediction of scores of individual questions within the MMSE questionnaire. In this setting, the concatenated visual and structured data embeddings are fed to multiple prediction

networks, each supervised by the score on a single question. This more granular design allows for the prediction of scores across various cognitive categories assessed by the MMSE, providing a detailed cognitive profile for each patient. Results are reported in Tab. 12.2, and confirm the general trend reported in Tab. 12.1, although the overall accuracy is lower, which can be explained by the increased difficulty of predicting individual question scores.

The standard deviations reported in Tabs. 12.1 and 12.2 help qualify the observed improvements by showing that the best configuration remains stable under repeated evaluations. At the same time, because the cohort is limited and project-specific, these values should be interpreted as indicators of experimental variability rather than as formal evidence of population-level statistical significance. A stronger inferential analysis would require larger cohorts, independent test sites, confidence intervals, and paired statistical tests designed for the adopted validation protocol.

Table 12.1: MMSE prediction accuracy using various combinations of task data (T), emotions (E) and visual pre-processing techniques, including face stabilization and cropping (FSC), video motion magnification (VMM), and MHI.

T	E	FSC	VMM	MHI	Sens.	Spec.	Acc.	F1
✓					38.4 ± 2.5	88.8 ± 1.7	63.6 ± 2.2	38.2 ± 2.4
	✓				47.0 ± 2.3	92.1 ± 1.4	69.6 ± 2.0	46.3 ± 2.2
✓	✓				49.9 ± 2.1	94.1 ± 1.2	72.0 ± 1.8	50.1 ± 2.1
		✓	✓	✓	72.8 ± 1.8	97.9 ± 0.9	85.4 ± 1.4	77.2 ± 1.6
✓	✓	✓		✓	75.4 ± 1.7	$\mathbf{98.1 \pm 0.8}$	86.8 ± 1.3	78.7 ± 1.5
✓	✓	✓	✓	✓	$\mathbf{82.6 \pm 1.5}$	97.9 ± 0.9	$\mathbf{90.3 \pm 1.1}$	$\mathbf{86.2 \pm 1.3}$

Table 12.2: Individual question prediction accuracy using various combinations of task data (T), emotions (E) and visual pre-processing techniques.

T	E	FSC	VMM	MHI	Sens.	Spec.	Acc.	F1
✓					32.1 ± 2.6	83.3 ± 1.9	57.7 ± 2.3	31.7 ± 2.5
	✓				53.8 ± 2.2	82.3 ± 2.0	68.1 ± 1.8	54.0 ± 2.1
✓	✓				73.7 ± 1.7	84.4 ± 1.8	79.1 ± 1.5	73.8 ± 1.6
		✓	✓	✓	74.8 ± 1.6	85.1 ± 1.5	80.0 ± 1.4	75.0 ± 1.5
✓	✓	✓		✓	75.1 ± 1.5	85.4 ± 1.4	80.3 ± 1.2	75.0 ± 1.4
✓	✓	✓	✓	✓	$\mathbf{75.2 \pm 1.4}$	$\mathbf{85.5 \pm 1.3}$	$\mathbf{80.4 \pm 1.1}$	$\mathbf{75.1 \pm 1.3}$

Finally, we investigate the impact of different temporal resolutions on the quality and utility of the generated MHIs. Temporal granularity plays a central role in how motion information is encapsulated within each frame, affecting the trade-off between information

compression and the expressiveness of the motion history. Tab. 12.3 shows performance metrics for different temporal lengths of a MHI, in seconds. A small history window results in lower accuracy, as it produces long input MHI sequences which may lead to excessive feature compression due to adaptive pooling; on the other hand, a large window seems to reduce the effectiveness of motion history, since recent motion may overwrite past information. Our choice of a 3-second window is justified by results as a good compromise.

Table 12.3: MMSE prediction accuracy at different MHI granularities.

MHI(sec)	**Sens.**	**Spec.**	**Acc.**	**F1**
1	80.0 ± 1.7	$\mathbf{98.1 \pm 0.8}$	89.1 ± 1.2	80.9 ± 1.5
2	77.5 ± 1.9	98.0 ± 0.9	87.8 ± 1.4	80.9 ± 1.7
3 (ours)	$\mathbf{82.6 \pm 1.4}$	97.9 ± 0.8	$\mathbf{90.3 \pm 1.0}$	$\mathbf{86.2 \pm 1.2}$
4	80.4 ± 1.6	98.0 ± 0.9	89.2 ± 1.2	83.0 ± 1.4
5	70.7 ± 2.4	97.5 ± 1.1	84.1 ± 1.8	73.7 ± 2.2

12.4 Discussion

Our results suggest that multimodal learning from tablet-based cognitive sessions can provide a feasible pathway toward scalable MMSE-related assessment, but they also clarify which components carry most of the practical signal and where caution is warranted. Across experiments, the strongest contribution consistently comes from vision-derived cues, indicating that facial behavior, in particular subtle dynamics that are typically suppressed in commodity recordings, contains discriminative information that can complement conventional task outcomes. In this sense, the benefit observed when introducing motion magnification is not merely an incremental accuracy improvement, but an operational finding: enhancing micro-movements can partially compensate for acquisition limitations (frame rate, noise, pose and illumination variability) that otherwise reduce the reliability of direct facial feature extraction. At the same time, the added value of task performance metrics and inferred affective states supports the interpretation that cognitive performance during structured activities is expressed through a joint behavioral signature rather than a single modality, and that contextual signals can refine predictions when facial evidence is ambiguous or noisy.

A limitation of the approach is that motion magnification remains sensitive to acquisition quality. Macro-movements, large head rotations, abrupt illumination variations, tracking errors, and sensor noise can introduce artifacts during magnification and may be propagated to the downstream model. The proposed pipeline mitigates this issue through face detection, face cropping, video stabilization, fine face re-cropping, and phase-based temporal band-pass

filtering, which emphasizes small motion components within a selected frequency range while reducing the influence of global drifts and out-of-band noise. However, these operations do not fully eliminate the problem: if global motion, lighting changes, or noise occur within the amplified frequency band, they can still be enhanced. Therefore, the present results should be interpreted as evidence of feasibility under controlled or semi-controlled acquisition conditions. Future work should include explicit video-quality assessment, illumination normalization, motion-outlier rejection, and confidence-aware abstention when the input is not suitable for reliable magnification.

The revised quantitative reporting includes standard deviations for the main MMSE and per-question prediction experiments. These values provide a useful measure of the stability of the observed trends: configurations including visual information, and especially motion magnification, remain consistently stronger than structured-only baselines under the adopted protocol. Nevertheless, given the limited cohort size and the project-specific acquisition setting, these results should not be interpreted as formal statistical significance claims. The present analysis supports methodological feasibility, while stronger statistical inference would require larger independent cohorts, confidence intervals, and paired tests comparing model variants under controlled validation schemes.

A second implication concerns clinical utility. Predicting an overall MMSE score is a useful first step, but cognitive screening decisions are often driven by patterns across subdomains. The model's extension to estimate per-question outcomes points toward a more granular representation of cognitive status, potentially enabling richer profiles that can guide follow-up testing or clinician review. However, this granularity also raises methodological requirements: per-item prediction is more sensitive to label noise, subject strategy, and task interface effects, and therefore needs careful validation to ensure that the inferred profile reflects cognitive factors rather than modality-specific artifacts or interaction confounds. Relatedly, the strong influence of visual features motivates explicit quality assessment and confidence reporting, because real-world recordings may include occlusions, atypical lighting, and non-cooperative behavior that can induce overconfident failures if not detected.

The study also highlights several directions for strengthening translation. Broader evaluation across acquisition devices and populations is necessary to quantify generalization and fairness, and longitudinal protocols are needed to test whether the learned behavioral signature is stable within subjects over time and sensitive to meaningful change. From a deployment perspective, edge-friendly inference and privacy-aware processing are also central, since facial video is identifiable and acceptance depends on minimizing retention and enabling on-device computation. Overall, these findings support the promise of facial micro-dynamics as a non-invasive marker within multimodal cognitive assessment pipelines, while emphasizing

the importance of robustness, calibration, and clinically grounded validation as the next steps toward practical monitoring tools.

12.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 340 hours of computation was performed on two NVIDIA A100 GPUs (thermal design power of 300 W), with total emissions estimated[1] to be 36.72 kg CO_2eq.

[1]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 13

Automated MoCA Score Estimation Using Eye-Gaze Data and Vision Transformers

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Isaak Kavasidis, Federica Proietto Salanitri, Lisa Passarello, Giovanni Piccininno, Vincenzo Masciale, Alessandro Anselmo, Cristiano Convertino, Domenico Rotondi, Nicola Laurieri, et al. Automated moca score estimation using eye-gaze data and vision transformers. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 963–968. IEEE, 2025

13.1 Overview

Cognitive impairment is a growing public health concern, and its early detection is a key factor for timely intervention and improved patient outcomes. MoCA [265] is widely adopted to screen MCI and early-stage dementia, yet its use in practice still depends on face-to-face administration and manual scoring by trained professionals. This creates a bottleneck that is simultaneously operational (time and personnel demands), methodological (susceptibility to human error and inter-rater variability), and practical (limited scalability for repeated assessments and longitudinal follow-up). These limitations motivate automated pipelines that can provide objective, repeatable estimates while reducing dependence on manual processing [182].

Among non-invasive signals, eye-gaze tracking is particularly promising because it captures subtle alterations in attention allocation, visual exploration, and executive control that may

precede overt performance deficits. In structured cognitive tasks, gaze trajectories encode both spatial preferences (which task regions attract attention) and temporal dynamics (how efficiently the subject scans, revisits, or fixates), thereby providing a rich but challenging data source for learning-based assessment. However, many existing approaches still rely on manual feature engineering and heuristic descriptors, which can be labor-intensive and introduce inconsistencies due to subjective design choices [27]. In addition, gaze acquisition in realistic settings often exhibits irregular sampling and missing points, complicating temporal modeling and reducing robustness.

To address these challenges, we propose an automated pipeline for MoCA score estimation that leverages pretrained ViTs for representation learning from eye-gaze data. Raw gaze samples are preprocessed and mapped onto task-relevant image regions, enabling a ViT encoder to extract high-dimensional representations that capture spatial structure in a way that is consistent across tasks and subjects. To explicitly handle gaze sampling irregularities and improve temporal reasoning, we introduce a time-aware positional embedding mechanism that encodes non-uniform temporal gaps, allowing the model to better interpret gaze dynamics beyond simple sequence order. The resulting representations are then processed by a transformer-based prediction module to estimate MoCA scores end-to-end, removing the need for handcrafted gaze features while preserving the temporal and spatial signatures that are informative for cognitive performance.

We validate the approach on a dataset collected during a structured protocol composed of seven cognitive gaming sessions, showing that transformer-based modeling of gaze patterns can provide a reliable and efficient alternative to manual MoCA processing. Overall, the study supports the feasibility of scalable, objective MoCA-related screening from eye-gaze signals, and positions time-aware transformer architectures as a practical foundation for automated cognitive assessment under realistic acquisition constraints.

13.2 Method

13.2.1 The proposed method

We propose a transformer-based architecture for automated cognitive assessment, which predicts MoCA scores using gaze-aligned visual input. The core idea is to process only the image regions where visual attention occurred, thereby reducing input dimensionality and enhancing focus on cognitively relevant content. To further improve temporal modeling, we introduce a novel Time-Aware Positional Embedding (TAPE) mechanism that accounts for non-uniform gaze sampling intervals and encodes temporal attention dynamics. The

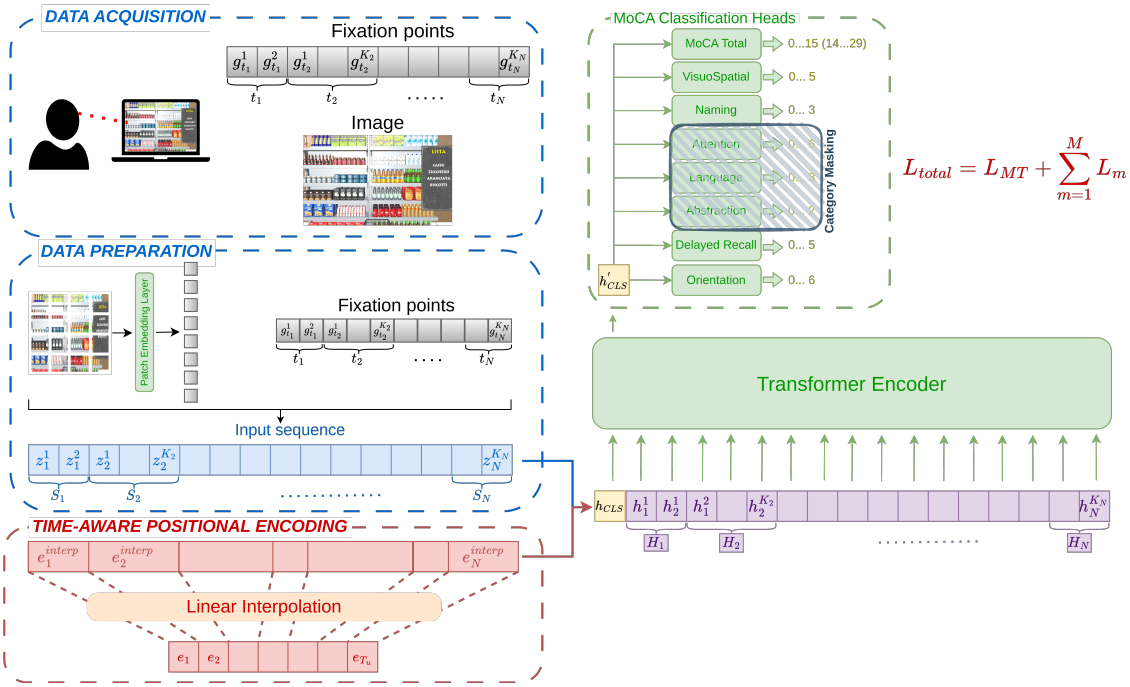


Figure 13.1: DA-ViT (Dynamic-Aligned Vision Transformer) architecture. Our major contribution lies in replacing the standard positional encoder with TAPE, which incorporates information about the temporal duration of fixation point. This helps standardize the representation of time, compensating for the highly variable eye-gaze sampling rates, which can differ significantly across consecutive time intervals.

architecture integrates a pretrained ViT to model both spatial and temporal dependencies in gaze-driven visual input and produces MoCA score predictions via multi-head classification.

13.2.2 Gaze-Conditioned Patch Selection

During data collection, participants interacted with a series of structured cognitive games displayed on a tablet device, while their eye-gaze coordinates $\mathbf{g}_t = (x_t, y_t)$ were continuously recorded using an eye tracker mounted directly on the tablet. Simultaneously, screenshots of the tablet interface were captured at a fixed frame rate and timestamped, ensuring temporal alignment between visual stimuli and gaze data.

To construct a compact and informative input sequence, we leverage gaze data to select only the embedded patches corresponding to regions actually attended by the participant. Specifically, we consider fixed-duration input segments of T seconds (e.g., $T = 10$), during which a single game frame $\mathbf{I}$ is displayed. Each frame $\mathbf{I}$ is divided into a non-overlapping grid of $P \times P$ patches, following the standard ViT tokenization scheme. All patches are independently passed through a shared patch embedding layer, producing a spatial token tensor $\mathbf{Z} \in \mathbb{R}^{P^2 \times d}$, where d is the embedding dimension.

Each input segment is temporally partitioned into $N = T/t_w$ non-overlapping windows of duration t_w seconds (e.g., $t_w = 1$), which are used to capture the temporal structure of gaze behavior. For each window $w = 1, \ldots, N$, we generate a gaze-driven sequence of patch tokens as follows:

$$\mathcal{S}_w = \left[\mathbf{z}_1^{(w)}, \mathbf{z}_2^{(w)}, \ldots, \mathbf{z}_{K_w}^{(w)}\right] \tag{13.1}$$

Here, $\mathbf{z}_k^{(w)}$ denotes the embedding of the k-th patch $\mathcal{P}_k \subseteq \mathbf{I}$ that is intersected by a gaze point at time $t_k \in [(w-1)t_w, wt_w]$, indicating that the gaze event occurred within the w-th temporal window. Thus, each token $\mathbf{z}_k^{(w)}$ represents a viewed patch, and the sequence $\mathcal{S}_w$ preserves the temporal order of the corresponding gaze events.

The window index w serves as a proxy for temporal position and is later used to generate **Time-Aware Positional Embeddings**, allowing the model to incorporate fine-grained temporal context in the encoding of gaze behavior.

This patch selection strategy ensures that only gaze-relevant regions are processed while retaining the spatial fidelity of visual stimuli. By computing patch embeddings independently of gaze data and selecting attended regions only afterward, the model preserves the full visual detail of each frame while reducing the computational cost of sequence processing. This allows attention to be focused on cognitively meaningful regions without sacrificing representational fidelity.

13.2.3 Time-Aware Positional Embedding (TAPE)

Eye-gaze data is often affected by inconsistencies due to both intrinsic factors (e.g., operating system scheduling, CPU load) and extrinsic factors (e.g., inaccurate gaze acquisition caused by improper face positioning or illumination changes). These issues result in variable sampling rates and irregular temporal distributions of gaze points across different sequences.

To incorporate temporal structure into gaze-driven input sequences, we introduce a **TAPE** mechanism that is robust to variable sampling rates in eye-tracking data.

In contrast to standard Vision Transformers, which use a fixed-size learnable positional embedding with length equal to the input sequence, our setting requires a more flexible formulation. Each input segment of duration T seconds is divided into T_u non-overlapping temporal windows of fixed duration t_w seconds (i.e., $T_u = T/t_w$). Each window may contain a variable number of gaze points and, consequently, a variable number of gaze-attended patch tokens.

To model temporal structure across windows, we learn a sequence of positional embeddings $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_{T_u}]$, where $\mathbf{e}_w \in \mathbb{R}^d$ encodes the temporal position of the w-th window. Since the number of attended patches in each window can vary across samples, we interpolate each

embedding vector $\mathbf{e}_w$ over the actual number of tokens n_w observed in window w.

Let $\mathbf{e}_w^{\text{interp}} \in \mathbb{R}^{n_w \times d}$ denote the interpolated version of $\mathbf{e}_w$, obtained via linear interpolation over n_w steps. For each patch token $\mathbf{z}_i^{(w)}$ observed in the time window w, the final temporally enriched token becomes:

$$\mathbf{h}_i^{(w)} = \mathbf{z}_i^{(w)} + \mathbf{e}_w^{\text{interp}}[i] \quad \text{for } i = 1, \ldots, n_w \tag{13.2}$$

This strategy allows each patch token to be augmented with a time-aware embedding that is both globally learnable (shared across sequences) and locally adaptable (interpolated to per-sample sequence length), enabling consistent modeling of temporal dynamics despite sampling variability.

The resulting temporally ordered sequence

$$\mathcal{H}_w = \left\{ \mathbf{h}_1^{(1)}, \ldots, \mathbf{h}_{n_1}^{(1)}, \ldots, \mathbf{h}_1^{(T_u)}, \ldots, \mathbf{h}_{n_{T_u}}^{(T_u)} \right\} \tag{13.3}$$

is then passed to the transformer encoder.

13.2.4 Transformer-Based Gaze Encoding

The temporally enriched sequence

$$\mathcal{H}_w = \{ \mathbf{h}_1^{(1)}, \ldots, \mathbf{h}_{n_1}^{(1)}, \ldots, \mathbf{h}_1^{(T_u)}, \ldots, \mathbf{h}_{n_{T_u}}^{(T_u)} \} \tag{13.4}$$

obtained through the TAPE mechanism, is passed to a ViT to model long-range dependencies across gaze-driven visual tokens.

A learnable classification token $\mathbf{h}_{\texttt{[CLS]}}$ is pre-pended to the sequence to summarize the global context:

$$\mathbf{H} = [\mathbf{h}_{\texttt{[CLS]}}, \mathbf{h}_1, \ldots, \mathbf{h}_L] \tag{13.5}$$

where $L = \sum_{w=1}^{T_u} n_w$ is the total number of attended tokens across all windows in the input segment.

This sequence is then processed through the transformer encoder, which applies multiple layers of self-attention and feedforward transformations:

$$\mathbf{H}' = \text{TransformerEncoder}(\mathbf{H}) \tag{13.6}$$

The output embedding corresponding to the classification token, denoted as $\mathbf{h}'_{\texttt{[CLS]}} \in \mathbb{R}^d$, serves as a compact representation of the participant's spatiotemporal gaze behavior over the

input segment.

This final embedding is then used for cognitive score prediction.

13.2.5 MoCA Score Prediction

We formulate MoCA estimation as a multi-head classification task, where each head predicts one of the MoCA categories as well as the total score. Let M be the number of MoCA categories, and let $\hat{\mathbf{y}}_m$ and $\hat{\mathbf{y}}_T$ denote the predicted probability distributions for category m and the total score, respectively.

Each prediction head consists of a fully connected layer followed by a softmax activation. The model is trained by minimizing the sum of categorical cross-entropy losses:

$$L_{\text{total}} = L_{MT} + \sum_{m=1}^{M} L_m \tag{13.7}$$

where L_{MT} is the categorical cross-entropy loss for the total MoCA score, and L_m is the loss for the m-th individual MoCA category. The cross-entropy loss for each L_m is defined as:

$$L_m = -\sum_{c=1}^{C} y_c^{(m)} \log\left(\hat{y}_c^{(m)}\right) \tag{13.8}$$

where C is the number of discrete classes for category m, $y_c^{(m)} \in 0, 1$ is the one-hot encoded ground-truth label, and $\hat{y}_c^{(m)}$ is the predicted probability for class c. Regarding the computation of the MoCA categories loss, the ground-truth labels for each game are multiplied by a binary mask that enables the model to back-propagate error gradients only to the category classification heads that the game-related data is associated with.

This approach ensures that only meaningful visual information is processed, reducing noise and improving model interpretability. By leveraging the pretrained ViT and integrating the time-aware embeddings, the method effectively captures the spatial and temporal patterns of visual attention, making it a powerful tool for cognitive assessment. The resulting model (Fig. 13.1) offers an automated alternative to manual MoCA scoring, potentially enabling more efficient and objective cognitive screening in clinical and research settings.

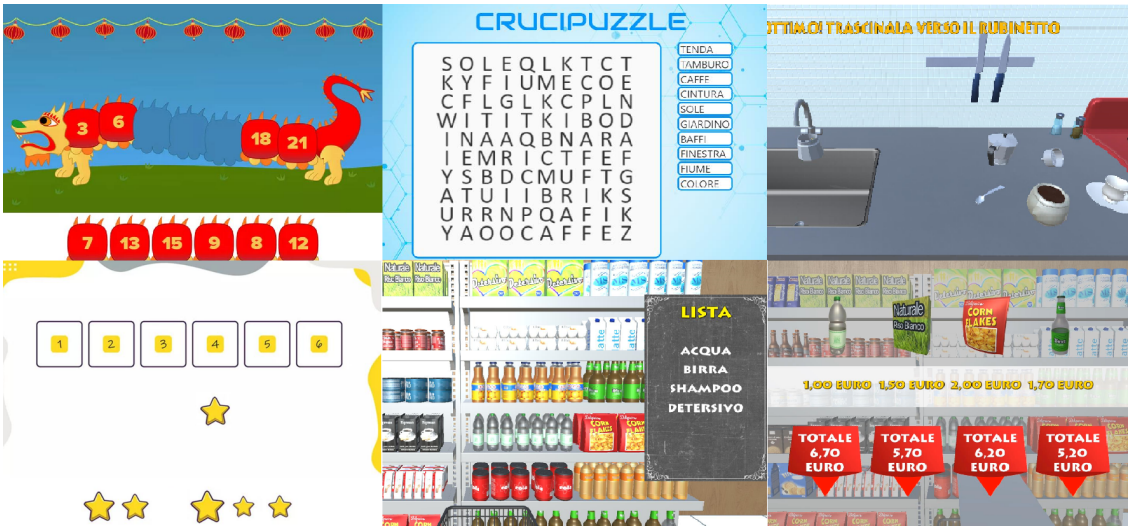


Figure 13.2: Example screenshots from the cognitive games presented to the subjects. From the top left: *Complete the sequence*, *Word search*, *Coffee preparation*, *Order by size*, *At the grocery shop*, and *Bill payment*.

13.3 Evaluation

13.3.1 Dataset acquisition

The study protocol involved the participation of 30 patients diagnosed with MCI, recruited from the Neuro-Geriatrics outpatient clinic at Cannizzaro Hospital in Catania. Each patient completed the MoCA questionnaire, with scores ranging from 14 to 29. The collected dataset includes a wide range of multi-modal data: eye-tracking data were recorded at a variable sample rate of 30-120 Hz, capturing saccadic movements, fixations, and smooth tracking. These parameters are relevant for assessing patients' attention and visual behavior during cognitive exercises. Using the eye-tracking data, saliency maps were generated to highlight which regions of the visual interface in the cognitive games attracted the participants' attention most intensely. Screen recordings documented patients' interactions with the cognitive games, allowing an in-depth analysis of their choices and responses during cognitive tasks, including response time, answer accuracy, and action sequence.

Since the proposed approach depends on gaze-to-task spatial mapping, gaze acquisition precision is an important practical factor. In this study, gaze samples were acquired after the standard calibration procedure of the eye-tracking device and were expressed in the screen coordinate system of the cognitive games. This allowed each gaze point or fixation to be associated with the corresponding interface layout and, subsequently, with task-relevant visual regions used by the gaze-conditioned representation. However, the dataset does not

Table 13.1: Game-MoCA attribute associations. VS: Visuo-spatial, N: Naming, AT: Attention, L: Language, AB: Abstraction, DR: Delayed recall, O: Orientation.

Game	**VS**	**N**	**AT**	**L**	**AB**	**DR**	**O**
Coffee preparation		✓				✓	✓
At the grocery shop		✓				✓	
Bill payment			✓				
Action and reaction		✓		✓		✓	
Word search			✓				
Order by size	✓		✓				
Complete the sequence	✓		✓		✓	✓	

include an independent device-level precision benchmark, repeated calibration-error estimates, or a cross-device validation protocol. Therefore, the spatial precision of the gaze signal cannot be reported as a device-independent quantity. Possible sources of mapping error include calibration drift, head-position changes, glasses reflections, illumination variations, and temporary tracking loss. These errors can affect the assignment of gaze samples to task regions or image patches and should be considered when interpreting the reported results.

The cognitive training consisted of seven digital games, designed to assess and stimulate different cognitive and functional abilities. Each game was also mapped to specific cognitive functions assessed by the MoCA. The *Word search* and *Order by size* were associated with visuo-spatial and executive functions, while *Coffee preparation*, *At the grocery shop*, and *Action and reaction* were linked to naming abilities. Attention was assessed through *Bill payment*, *Word search*, *Order by size*, and *Complete the sequence*, while *Action and reaction* also engaged language skills. Abstract reasoning was stimulated by *Complete the sequence*, and delayed recall was targeted in *Coffee preparation*, *At the grocery shop*, *Action and reaction*, and *Complete the sequence*. Finally, *Coffee preparation* also contributed to orientation assessment. Table 13.1 reports the games and the associated MoCA attributes, while Fig. 13.2 shows example screenshots of the games.

Each patient participated in a single recording session lasting approximately one hour. Additionally, all participants signed an informed consent form, which included a detailed section on data collection and usage. Along with the multi-modal data, demographic information such as age, sex, and educational level was collected, along with clinical details including medical history and current medications. The study included 13 male and 17 female participants, with ages ranging from 59 to 87 years (mean age: 75.4). Educational levels varied among participants: 2 had no formal education, 5 completed elementary school, 6 attended lower secondary school, 10 completed upper secondary school, and 7 had a university degree. Cognitive and functional assessments, including previous evaluations and MoCA

Table 13.2: Performance comparison for different MoCA-related tasks.

Task	**Balanced accuracy(%)**
Total MoCA Classification	83.52 ± 0.10
MoCA Category Classification	87.51 ± 0.69
Total MoCA + MoCA Category (Ours)	**89.93 ± 0.12**

scores, were also documented to provide a comprehensive overview of each patient's condition.

13.3.2 Experimental setting

For all experiments, the model was trained for 100 epochs using the AdamW optimizer [218] with a weight decay of 5×10^{-5}. As backbone, we used the *ViT-B/16* Vision Transformer pretrained on ImageNet. The learning rate was set to $1e-4$ and the batch size was set to 16. The input images were resized from a 1920×1020 initial resolution to 224×224, which is the suitable input resolution for ViT patch embeddings. Due to the relatively small number of data samples ($N = 309$), *5-fold* cross-validation ($N_{Fold} = 62$, except the last one that has 61 items) was used and the obtained performance is expressed in terms of average values and standard deviations across all folds.

13.3.3 Performance

We evaluated the performance of the proposed pipeline on the dataset presented. Table 13.2 reports the model's accuracy in predicting both the total MoCA score and the individual category scores. To highlight the importance of formulating the task as a classification problem at both levels, we report the results for the total and category-level predictions separately within the same table.

The results show that training the model with both the total MoCA score and individual category scores (89.93%) leads to better performance than training it solely on the total MoCA score (83.52%), with an improvement of nearly 6.5 percentage points. It also outperforms the model trained only on individual MoCA categories (87.51%) by about 1.5 percentage points. This suggests that while the primary contribution to classification performance comes from the individual MoCA categories, incorporating the total MoCA score improves performance.

The standard deviations reported in Tab. 13.2 summarize variability across the cross-validation protocol. They indicate that the combined total-score and category-level formulation is stable across folds under the adopted dataset split. However, because the cohort size remains limited, these values should be interpreted as fold-level variability rather than as definitive confidence intervals on population-level performance.

Table 13.3: Performance comparison when different positional embedding methods are used.

Positional Embedding	**Balanced accuracy(%)**
FIXED	79.45 ± 0.36
LEARNABLE	85.72 ± 2.10
TAPE (Ours)	**89.93 ± 0.12**

Table 13.4: Performance comparison using different time unit durations.

t_w	**Balanced accuracy(%)**
0.5 sec	89.57 ± 0.17
1 sec (Ours)	**89.93 ± 0.12**
2 sec	85.72 ± 0.66
5 sec	84.18 ± 0.65

In order to evaluate the impact of the time-aware positional embedding, we also conducted the same training experiments by replacing it with a learnable linear positional encoding [85] and a vanilla transformer (sine/cosine) positional embedding [373]. The results, shown in Table 13.3, demonstrate that the model with the time-aware positional embedding outperforms, on average by several percentage points, the learnable and fixed positional embeddings.

Additionally, in Table 13.4, we report the results obtained by varying the granularity of the time unit length (t_w). The results indicate that the smaller the time unit (i.e., the shorter the temporal window of eye-gaze data it contains), the more accurate the model becomes. However, reducing the time-unit length below 1 second does not provide a meaningful improvement in accuracy while increasing the model's complexity, leading to premature model overfitting.

The ablation results in Tabs. 13.3 and 13.4 also include standard deviations across folds, allowing the stability of the proposed TAPE configuration to be assessed relative to alternative positional encodings and temporal granularities. The comparison supports the empirical robustness of the proposed design under the available data; nevertheless, the limited sample size prevents a full statistical significance analysis. Larger cohorts would be required to compute reliable confidence intervals and to perform paired tests between model variants.

13.4 Discussion

This work investigates automated MoCA score estimation from eye-gaze trajectories by combining gaze-to-task spatial mapping with pretrained Vision Transformers. The results support the feasibility of learning clinically relevant signals from visual-attention dynamics acquired during structured cognitive tasks, and suggest that transformer-based representations

can reduce dependence on manual gaze feature engineering. In particular, the proposed time-aware positional embedding addresses a practical limitation of real gaze recordings (inconsistent sampling and irregular temporal gaps) and provides a principled way to preserve temporal information that would otherwise be distorted by naive resampling or uniform-position assumptions. Beyond performance improvements, this design choice has an important operational implication: it moves the pipeline closer to realistic acquisition conditions, where missing samples and variable sampling rates are the norm rather than an exception.

From a statistical perspective, the reported mean and standard deviation values provide evidence that the proposed configuration is stable across the adopted cross-validation protocol, but they should not be interpreted as definitive statistical significance claims. The dataset size limits the reliability of formal inferential testing, particularly for paired comparisons across model variants and for estimating confidence intervals that generalize to broader populations. For this reason, the results are interpreted as methodological evidence supporting feasibility, while stronger conclusions will require larger cohorts, independent validation, and pre-specified statistical analyses.

At the same time, several considerations temper the interpretation of the current findings. First, validation was conducted on a limited dataset acquired from cognitive gaming sessions, and the resulting performance should be read as evidence of promise rather than as a definitive estimate of clinical generalization. Eye-gaze behavior is sensitive to task design, screen layout, calibration quality, device characteristics, and participant-specific factors (e.g., familiarity with tablets, visual acuity, fatigue), all of which can introduce distribution shift. Consequently, the next methodological step is to quantify robustness under cross-session, cross-device, and cross-population splits, reporting not only aggregate accuracy but also calibration and error patterns near clinically meaningful thresholds. Second, although the pipeline is designed to be automatic, it still depends on reliable gaze acquisition and on accurate mapping to task-relevant regions. Failures in gaze tracking (e.g., slippage, illumination changes, glasses reflections) can degrade representations in ways that resemble cognitive impairment. This motivates explicit quality assessment and conservative fallback logic, so that low-quality acquisition leads to abstention or re-acquisition rather than overconfident scoring.

The dependence on gaze-to-task spatial mapping also limits direct generalization to different eye-tracking devices and data collection protocols. A model trained on one screen layout, calibration procedure, sampling rate, and coordinate system may not transfer directly to another setup, because the same visual-attention behavior can produce different gaze distributions when interface geometry, screen size, viewing distance, or device precision changes. Generalization to alternative protocols would therefore require subject-level recalibration before acquisition, spatial standardization of screen coordinates and task regions,

quality-aware filtering of low-confidence gaze samples, and preferably training on multi-device or multi-protocol datasets. Another promising direction is to encode gaze in normalized task-relative coordinates rather than only in absolute screen coordinates, so that fixations are represented with respect to semantic regions of the cognitive game. This would reduce dependence on a specific acquisition setup and support more reliable deployment across clinics, tablets, and eye-tracking hardware.

A further limitation is that MoCA is a composite assessment: the final score aggregates heterogeneous cognitive skills, and different behavioral strategies can lead to similar outcomes. Eye-gaze captures attention allocation and visual exploration, but it may under-represent other latent factors that influence MoCA performance. This motivates a clinically meaningful role for multimodal fusion when acceptable and consented. Combining gaze with complementary signals such as facial behavior or speech could broaden coverage of cognitive dimensions and reduce ambiguity, particularly in borderline cases. Along this line, integrating non-invasive behavioral modalities within a shared temporal model is a natural extension, consistent with evidence that vision-based micro-dynamics can be informative for cognitive scoring [246]. More generally, multimodal fusion should be designed to remain functional under missing channels, so that gaze-only inference remains the deployable default while additional signals serve as optional context.

From a deployment perspective, these findings indicate a plausible pathway toward scalable cognitive screening: a low-burden, repeatable, and objective assessment that reduces manual scoring effort and mitigates human error. However, translation requires clinically grounded validation. Prospective studies should compare model outputs against clinician-administered MoCA across repeated visits, quantify longitudinal consistency, and assess whether the system improves screening throughput or supports earlier referral decisions. Equally important are human factors: interfaces should present uncertainty-aware outputs and actionable summaries rather than a single score, ensuring that the tool supports decision making without overstating confidence.

Overall, this study positions gaze-based transformer modeling as a promising foundation for automated MoCA estimation under realistic acquisition constraints. The key research directions are clear: larger and more diverse datasets, explicit quality and calibration analysis, robustness under device and population shift, and multimodal extensions that preserve privacy and deployability while improving clinical relevance.

13.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 315 hours of computation was performed on an NVIDIA A100 GPU (thermal design power of 300 W), with total emissions estimated[1] to be 34.02 kg CO_2eq.

[1]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 14

Ensemble and personalized transformer models for subject identification and relapse detection

Source note. The following chapter is based on and adapted from:

- Salvatore Calcagno, Raffaele Mineo, Daniela Giordano, and Concetto Spampinato. Ensemble and personalized transformer models for subject identification and relapse detection in e-prevention challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023

14.1 Introduction

Long-term passive sensing through consumer wearables is increasingly explored as a practical route to quantify behavioral and physiological patterns in everyday life, with the aim of supporting longitudinal monitoring and risk assessment in psychiatric conditions. The e-Prevention project introduced an integrated framework for multimodal acquisition and analysis, combining smartwatch-derived biosignals (e.g., inertial streams and cardiac markers) with clinical context, to enable data-driven monitoring and relapse prevention in psychotic disorders [427]. Within this context, the ICASSP 2023 e-Prevention Signal Processing Grand Challenge formalized two complementary tasks on long-term wearable time series acquired in unconstrained daily-life conditions, characterized by heterogeneous sampling, missing segments, and strong inter-subject variability. This chapter summarizes the PeRCeiVe Lab submissions for (i) closed-set subject identification from continuous smartwatch measurements (46 subjects) and (ii) relapse detection at day level for patients in the psychotic spectrum, a

setting where training data include only non-relapse days and relapse labels are available only for validation.

For Track 1, we design an ensemble that combines a deep 1D convolutional backbone for local temporal motif extraction with multiple transformer variants that differ in positional modeling (sinusoidal vs Time2Vec) and architectural width, with logit-level fusion to stabilize predictions under missingness and day-to-day variability. For Track 2, we adopt a personalized anomaly-detection formulation by training one autoencoder per subject using non-relapse data only and converting reconstruction errors into anomaly probabilities via empirical cumulative distribution functions, followed by robust day-level aggregation. This framing is consistent with prior results showing that deep models on short wearable segments can support reliable identification and that autoencoder-based formulations can capture deviations relevant to mental-health monitoring [312, 280].

On the official hidden test sets, our submissions achieved competitive performance in both tracks, ranking 2nd in subject identification with 93.85% accuracy and 1st in relapse detection with ROC-AUC 0.6469 and PR-AUC 0.6509 (harmonic mean 0.6489). Beyond the leaderboard outcome, the chapter highlights two practical design principles for longitudinal wearable analytics: leveraging architectural diversity and inference-time ensembling to improve robustness under real-world acquisition artifacts, and using patient-specific normality modeling to mitigate inter-subject variability when clinically meaningful labels are sparse.

14.2 Methods

Experiments are conducted on the wearable data released in the e-Prevention Signal Processing Grand Challenge (ICASSP 2023), derived from the e-Prevention project [427]. The acquisition platform is a consumer smartwatch (Samsung Gear S3 Frontier) providing inertial and cardiac-related signals, together with activity metadata such as steps and sleep scheduling [427]. The challenge defines two tracks with fixed train/validation/test splits. Labels are available only for train and validation, while test labels are kept hidden and used for the official ranking.

14.2.1 Track 1 - Ensemble for subject identification

The goal is closed-set identification of the smartwatch wearer among 46 subjects, using multivariate time series recorded over multiple months per subject (mean recording length reported by the organizers is approximately 2.5 months) [427]. The submitted system operates on windowed segments extracted from the continuous recordings.

The main submission uses an 8-channel subset consisting of 3-axis accelerometer, 3-axis

Table 14.1: Aggregated feature set for Track 2 (computed on 5-minute slices and stacked over the day) [427].

Group	**Features**
Motion (IMU)	Mean energy / mean norm of linear acceleration (accelerometer), mean energy / mean norm of angular acceleration (gyroscope)
Cardiac (PPG-derived)	Mean heart rate, mean RR interval
HRV descriptors (NNI)	Poincaré plot major axis (SD2), normalized LF power (Lomb-Scargle), normalized HF power (Lomb-Scargle)
Temporal context	Daily sinusoidal encoding: $\sin(\cdot)$ and $\cos(\cdot)$
Data quality	Percentage of valid samples (percentage of correctly identified pulses)

gyroscope, heart rate, and RR intervals. Each channel is min-max normalized to $[0, 1]$ using statistics computed on the training split. Missing samples inside windows are imputed by nearest-neighbor interpolation. During training, we retain only windows with at least 83% valid samples. Data augmentation is performed through segmentation into non-overlapping windows of fixed width. We employ two window sizes, corresponding to $T = 2160$ samples (3h) and $T = 1080$ samples (1.5h), consistent with the configurations used in the submitted system. Each window is represented as a matrix $\mathbf{X} \in \mathbb{R}^{F\times T}$, where F denotes the number of input channels.

All ensemble members share a convolutional front-end that maps $\mathbf{X} \in \mathbb{R}^{F\times T}$ to a latent sequence $\mathbf{Z} \in \mathbb{R}^{d\times L}$ through five 1D convolutional blocks (Conv1D, ReLU, BatchNorm, Dropout) followed by AdaptiveAvgPool1d to a fixed length L. This module acts as an embedding layer for transformer-based variants and also supports a CNN-only classifier baseline.

For transformer members, we prepend a learnable $[CLS]$ token and add a positional embedding. We consider both sinusoidal encoding [373] and Time2Vec [178]. The embedded sequence is processed by a Transformer Encoder with n_{layers} layers, n_{head} attention heads, model width d_{model}, and feed-forward dimension d_{hid}. The final $[CLS]$ representation is mapped to logits through a two-layer MLP head.

Let $\boldsymbol{\ell}_m(\mathbf{X}) \in \mathbb{R}^{46}$ be the logits produced by model m. The ensemble prediction is obtained by a weighted sum of logits,

$$\boldsymbol{\ell}(\mathbf{X}) = \sum_{m=1}^{M} w_m \, \boldsymbol{\ell}_m(\mathbf{X}), \tag{14.1}$$

followed by $\text{softmax}(\boldsymbol{\ell}(\mathbf{X}))$. The weighting accounts for duplicated configurations used with

Table 14.2: Track 1 ensemble configurations used in the submission. "PE" denotes positional embedding. A duplicated configuration is assigned weight 2 in the ensemble.

ID	T	Model	PE	d_{model}	n_{head}	d_{hid}	Weight
1	2160	Transformer	sin-cos	32	2	128	2
2	2160	Transformer	Time2Vec [178]	32	2	768	1
3	1080	Transformer	Time2Vec [178]	128	2	768	1
4	2160	Transformer	Time2Vec [178]	128	2	512	1
5	2160	CNN-only (ablation)	Time2Vec [178]	128	-	-	1

higher voting weight in the submission.

Each member is trained independently using cross-entropy loss and Adam optimizer with batch size 64. A ReduceLROnPlateau scheduler is applied with factor 0.5 and patience 10 epochs. Models are trained up to 300 epochs, selecting the final checkpoint based on validation accuracy.

14.2.2 Track 2 - Personalized relapse detection via reconstruction-based anomaly scoring

The goal is to infer, for each day of recording, whether the subject is in relapse or non-relapse. The released Track 2 subset involves 10 subjects after preprocessing [427]. Training data contain only non-relapse days, while relapse days are provided in validation; test labels are hidden. This structure motivates a personalized formulation where each subject is modeled independently. We use the challenge-provided aligned time-series release.

We consider both *raw* and *aggregated* representations. Raw data follow the same preprocessing used in Track 1, with subject-wise normalization and windowing. Aggregated data are computed from 5-minute slices to obtain a compact daily multivariate sequence, following the feature definitions provided in the dataset description [427]. Aggregated features are standardized with subject-wise mean and standard deviation estimated on the training split.

For each subject s, we train a subject-specific autoencoder f_s using only non-relapse windows from the training split. At inference time, relapse likelihood is inferred as deviation from the learned reconstruction distribution, mitigating inter-subject variability and label sparsity.

The personalized formulation is also intended to reduce false positives caused by benign physiological variability. Smartwatch-derived inertial and cardiac-related signals are non-specific: physical exertion, acute stress, sleep disruption, temporary sensor displacement, missing samples, or sensor noise can produce deviations that may resemble abnormal events if interpreted through global thresholds. Instead, the proposed approach models each subject

independently and evaluates whether the current day departs from that subject's own non-relapse baseline. In this way, the alert score is not driven by an absolute heart-rate or motion value alone, but by its deviation from the individual's learned normality pattern.

We consider three families of autoencoders: (i) a CNN1D autoencoder operating on raw or aggregated sequences, (ii) a Transformer Encoder-Decoder autoencoder operating on aggregated sequences, and (iii) an alternative time-series autoencoder implemented in the VolUnD toolkit[1]. For the CNN and Transformer models, we optionally enable a variational formulation by splitting the bottleneck into $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$ and adding a KL term to the reconstruction loss.

A grid search is executed independently per subject over learning rate, representation type (raw/aggregated when applicable), variational flag, and architectural hyperparameters. The selected configuration maximizes validation performance for that subject.

Models are trained for 100 epochs with MSE reconstruction loss (plus KL regularization if variational), using Adam optimizer and batch size 64. ReduceLROnPlateau is used with factor 0.5 and patience 10 epochs.

For a window $\mathbf{X}$ and its reconstruction $\widehat{\mathbf{X}} = f_s(\mathbf{X})$, we compute a per-channel reconstruction error vector $\mathbf{e}(\mathbf{X})$. From training non-relapse windows, we estimate for each channel the empirical cumulative distribution function $CDF(\cdot) = \mathbb{P}(E \leq x)$. At inference, the anomaly score is computed as

$$\alpha(\mathbf{X}) = CDF(\mathbf{e}(\mathbf{X})), \tag{14.2}$$

where CDF is applied channel-wise and aggregated by a robust statistic. Since a day contains a variable number of windows, we report the daily anomaly score as the median over all windows available for that day, and use it as the relapse probability proxy.

This window-to-day aggregation further mitigates isolated false positives: a single noisy segment, a transient exertion episode, or a short sensor artifact has less influence than a persistent deviation across the day. In addition, the aggregated feature set includes the percentage of valid samples, which provides a basic quality-related descriptor. However, this mechanism is not equivalent to a complete signal-quality module. It reduces sensitivity to isolated artifacts but does not fully distinguish relapse-related deviations from benign contextual changes, motivating future extensions with explicit activity labels, signal-quality indicators, and uncertainty-aware alert policies.

[1] https://github.com/EUROVOLC-ML/VolUnD-Toolkit

Table 14.3: Track 2 - Best per-subject configurations selected by grid search. "Aggr." indicates aggregated representation, "Raw" indicates raw representation.

Subject	Model	Data	d_{model}	n_{head}	LR
0	Transformer AE	Aggr.	32	8	$5 \cdot 10^{-3}$
1	Transformer AE	Aggr.	128	16	$5 \cdot 10^{-4}$
2	CNN AE	Aggr.	-	-	$5 \cdot 10^{-3}$
3	Transformer AE	Aggr.	32	4	$5 \cdot 10^{-4}$
4	Transformer AE	Aggr.	32	8	$5 \cdot 10^{-3}$
5	VolUnD AE	Raw	-	-	$5 \cdot 10^{-3}$
6	CNN AE	Raw	-	-	$5 \cdot 10^{-3}$
7	Transformer AE	Aggr.	128	8	$5 \cdot 10^{-3}$
8	Transformer AE	Aggr.	128	8	$5 \cdot 10^{-3}$
9	Transformer AE	Aggr.	128	8	$5 \cdot 10^{-3}$

14.3 Results

The proposed models were evaluated by the challenge organizers on the hidden test splits. For Track 1 (subject identification), performance is measured using raw classification accuracy. For Track 2 (relapse detection), the official score is defined as the harmonic mean between ROC-AUC and PR-AUC, to jointly account for ranking quality under class imbalance and threshold sensitivity. The official leaderboard results are reported in Table 14.4 (Track 1) and Table 14.5 (Track 2). Our submission achieved 93.85% test accuracy for Track 1, ranking 2nd among participating teams, and obtained ROC-AUC 0.6469 and PR-AUC 0.6509 for Track 2, corresponding to a harmonic mean of 0.6489 and ranking 1st.

Table 14.4: Performance comparison among participant teams on the provided test set for subject identification. We report the raw test accuracy for each team. Our score ranks 2nd.

Team	**Accuracy**
SRCB-LUL	95.00%
PeRCeiVe	**93.85%**
AI_Bezzie	91.36%
SAILers	82.15%
unipi_cmbl	75.43%
ADCADD	3.83%
ID-EPRE2	2.88%
eHust	2.88%
CogBCI	2.88%
uoi	2.68%
NWPU	2.3%

Table 14.5: Performance comparison among participant teams on the provided test set for relapse detection. We report the ROC-AUC, the PR-AUC and the harmonic mean of the previous scores, for each team. Our score ranks 1st.

Team	ROC-AUC	PR-AUC	Total
PeRCeiVe	**0.6469**	**0.6509**	**0.6489**
Emotion	0.6072	0.6347	0.6209
SAILers	0.5839	0.6263	0.6051
SmartBCI	0.5435	0.5863	0.5604
YDH@HEU	0.5215	0.5587	0.5401
GIPS@HEU	0.5117	0.5480	0.5229

To provide additional insight into the personalized relapse-detection pipeline, Table 14.6 reports the validation performance for each subject and model family, using the harmonic mean between ROC-AUC and PR-AUC. Results highlight a marked inter-subject variability and confirm that the best-performing choice depends on the subject and the data representation (raw versus aggregated), supporting the adopted subject-wise model selection strategy.

Table 14.6: Performance of tested models on validation set for each subject. We reported the harmonic mean between ROC-AUC and PR-AUC. For CNN and VolUnD we tested also *raw* and *aggregated* data.

model		subject									
		0	1	2	3	4	5	6	7	8	9
CNN	*raw*	.53	.66	.45	.73	.62	.57	**.74**	.70	.58	.62
	aggr.	.44	.75	**.51**	.82	.73	.58	.15	.84	.78	.76
VolUnD	*raw*	.58	.67	.49	.72	.59	**.64**	.57	.58	.64	.66
	aggr.	.42	.77	.46	.78	.82	.53	.14	.81	.76	.76
Transf.	*aggr.*	**.61**	**.84**	.51	**.83**	**1.00**	.61	.47	**.89**	**.94**	**.79**

During method development, we carried out additional validation analyses to contextualize the design choices behind the final submissions. An input-feature ablation (Table 14.7) evaluates the impact of including non-periodic channels (sleep and step-derived information) on a CNN baseline, showing that these signals can provide modest improvements in validation accuracy, while their benefit depends on the specific combination of channels. In addition, a window-length study (Table 14.8) indicates that moderate temporal context (1.5h-3h) yields a substantially better trade-off than much longer windows (15h), which tend to overfit and generalize poorly.

For Track 2, we additionally summarize VolUnD hyperparameter exploration results in Table 14.9, reporting the mean and standard deviation of the maximum validation score across runs, and the impact of retaining the best-performing configuration.

Table 14.7: Track 1 - Input-feature ablation on the CNN baseline (validation). "Sleep" indicates inclusion of sleeping status; "Steps" indicates step-derived signals; "Calories" indicates calories per time unit

Sleep	Steps	Calories	Max val accuracy
1	0	0	0.8768
1	1	0	0.8747
0	0	0	0.8687
0	1	1	0.8646
0	1	0	0.8606
1	1	1	0.8586

Table 14.8: Track 1 - Window size study (validation). Reported values are the final train and validation losses and accuracies for each window length

Window	Val acc
15h	0.7960
3h	0.9273
1.5h	0.9434

14.4 Discussion

This chapter reported the PeRCeiVe Lab submissions to the ICASSP 2023 e-Prevention Signal Processing Grand Challenge, addressing (i) subject identification from long-term smartwatch recordings and (ii) relapse detection in the psychotic spectrum. Across the two tracks, the adopted strategies reflect two complementary principles for longitudinal wearable analytics: architectural diversity to stabilize identification under real-world noise and missingness, and personalization to mitigate inter-subject variability when clinically relevant labels are sparse.

In Track 1, the 46-way subject identification task is characterized by substantial day-to-day variability, heterogeneous activity patterns, and missing segments typical of unconstrained wearable acquisition. In this setting, no single architecture provides uniformly optimal inductive bias across all temporal contexts. The proposed ensemble combines a convolutional front-end that captures local temporal motifs with Transformer variants that model longer-range dependencies after temporal pooling. The strong test performance (93.85%, 2nd overall) supports the practical benefit of fusing complementary representations at the logit level, which reduces sensitivity to a particular window configuration or positional embedding choice. The additional validation studies further clarify these design choices. First, the window-size analysis indicates that moderate temporal context (1.5h-3h) provides a favorable trade-off, while excessively long windows (e.g., 15h) increase optimization difficulty and

Table 14.9: Track 2 - VolUnD validation summary over the hyperparameter search (maximum validation score across runs).

Selection rule	**Mean**	**Std**
All evaluated configurations	0.7369	0.1105
Best configuration retained	0.7546	0.1026

may amplify the effect of missingness and distribution shift across days. Second, positional embedding ablations show comparable validation performance between sinusoidal encodings and Time2Vec, suggesting that, under the employed convolutional embedding and pooling, positional modeling is less of a bottleneck than robustness to activity variability and missing data. Third, fusion-scheme comparisons indicate that simple aggregation strategies (sum or mean of logits) are sufficient to exploit model diversity, without requiring more complex calibration layers.

Track 2 naturally motivates a personalized formulation. After preprocessing, only 10 subjects are considered, and the training split contains exclusively non-relapse days, while relapse labels appear only in validation [427]. Under these constraints, discriminative supervised learning is limited by label scarcity and by the strong heterogeneity of patient-specific baselines. Modeling relapse as a deviation from a learned subject-specific normality distribution is therefore a principled approach. The achieved top test ranking (harmonic mean 0.6489, 1st overall) indicates that reconstruction-based scoring can be competitive even when the absolute ROC-AUC and PR-AUC values remain moderate, reflecting the intrinsic difficulty of inferring relapse from noisy passive sensing. A key observation emerging from the per-subject validation analysis is that the best-performing architecture and representation depend on the subject. Transformer autoencoders operating on aggregated daily features often provide strong results, likely because aggregation suppresses short-term noise and enforces a consistent temporal grid, enabling attention mechanisms to capture structured daily dynamics. However, for some subjects, raw-signal models (CNN-based or alternative autoencoders) are preferable, suggesting that aggregation may remove subtle patterns informative for certain individuals. These findings reinforce the rationale for subject-wise model selection, and they motivate future work on meta-learning or conditional model selection mechanisms that retain personalization while improving operational simplicity. The proposed CDF-based scoring further contributes to robustness by mapping reconstruction errors to empirical probabilities derived from non-relapse training windows. This reduces sensitivity to the absolute scale of reconstruction error, which can vary across channels, architectures, and subjects, and provides a simple calibration mechanism compatible with window-to-day aggregation via the median statistic.

A specific limitation of Track 2 is the non-specificity of wearable physiological and inertial signals. Benign events such as high physical exertion, acute stress, irregular sleep, temporary device displacement, missing segments, and sensor noise can increase reconstruction error even when no relapse is present, potentially producing false positive relapse alerts. The proposed personalized design mitigates this risk because each model learns the normal distribution of an individual subject rather than relying on population-level thresholds. In addition, daily aggregation by the median reduces the influence of isolated noisy windows. Nevertheless, these mechanisms cannot fully separate clinically relevant relapse-related deviations from benign contextual changes. A more deployment-oriented system should therefore incorporate explicit signal-quality indicators, activity/context recognition from IMU and smartwatch metadata, missingness-aware uncertainty estimates, and conservative alert policies that require persistent deviations before escalation.

Several limitations should be considered when interpreting these results. First, Track 2 evaluation relies on a limited number of subjects and on an imbalanced distribution of relapse days, which constrains statistical confidence and may amplify sensitivity to split-specific artifacts. Second, the use of fixed window thresholds for validity and interpolation strategies may influence downstream performance; while these choices are necessary in practice, they are not uniquely determined and may interact with activity-dependent missingness. Third, the reported Track 2 performance is achieved through a personalized pipeline requiring training and maintaining one model per subject, which is appropriate for longitudinal monitoring but raises deployment questions, including model update policies, computational budgets, and handling of new subjects with limited baseline data. Fourth, the current pipeline includes basic data-validity information but does not yet implement a full quality-aware alerting layer able to distinguish sensor artifacts, benign exertion, and clinically meaningful deterioration in a systematic way.

Future extensions should explore unified formulations that preserve the benefits of ensembling and personalization while improving operational efficiency. For Track 1, this includes distillation of the ensemble into a single deployable model and explicit calibration to support confidence-aware identification under missingness. For Track 2, promising directions include hybrid objectives that combine reconstruction with self-supervised representation learning, subject-adaptive thresholds and temporal smoothing strategies, and uncertainty-aware scoring that accounts for varying data availability across days. More broadly, the challenge outcomes suggest that robust wearable analytics benefits from combining strong sequence models with careful handling of real-world acquisition constraints and patient-specific variability. Additional future work should also model benign contextual factors explicitly, so that activity, sleep, missingness, and signal-quality patterns can be used to reduce false positive relapse

alerts without suppressing persistent deviations that may require clinical attention.

14.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 190 hours of computation was performed on an NVIDIA RTX A6000 GPU (thermal design power of 300 W), with total emissions estimated[2] to be 24.62 kg CO_2eq.

[2]Estimations were performed using the Machine Learning Impact calculator [189].

Part III

Sign Language Recognition for Medical Communication

Chapter 15

Importance and Challenges in Automatic SLR

Effective communication between patients and clinicians is a prerequisite for accurate diagnosis, informed consent, adherence to therapy, and patient safety. For Deaf and hard-of-hearing individuals who primarily rely on sign language, communication barriers can translate into misunderstandings, reduced access to services, and inequities in care. Professional interpreting services remain the appropriate and preferred solution in many settings, yet availability is not guaranteed across hospitals, emergency departments, and smaller clinical facilities. Moreover, time-critical scenarios may demand immediate interaction when an interpreter is not readily accessible. In this context, sign language recognition (SLR) technologies can serve as assistive tools that improve accessibility, for example by supporting triage, basic symptom reporting, or structured medical interactions. The aim is not to replace interpreters, but to provide an additional pathway that reduces barriers when resources or timing constraints limit access to human mediation.

Medical sign language communication differs in important ways from generic SLR settings. First, clinical exchanges involve domain-specific vocabulary and phrasing, including symptom descriptions, anatomical references, medication terms, and procedure-related language. Second, the stakes are higher and tolerance for errors is lower: misrecognition can directly affect safety, which makes reliability, calibration, and explicit uncertainty handling essential rather than optional. Third, signing is heterogeneous in ways that become more pronounced in healthcare: regional variants, signer proficiency, stress-induced changes in speed and articulation, and coarticulation effects can all increase variability precisely when communication accuracy matters most. Medical dialogue is also strongly context-dependent: many signs and expressions require conversational context for disambiguation, and clinically meaningful interpretation often depends on sequential information rather than isolated recognition. These

characteristics motivate dedicated datasets and evaluation protocols tailored to healthcare communication, rather than direct transfer from generic benchmarks.

Even outside the medical domain, SLR remains technically challenging because sign languages are complex visual-gestural systems. Many signs differ by subtle, fine-grained variations in handshape, finger articulation, orientation, or motion trajectory, which are difficult to capture reliably with standard camera setups. Signing is inherently multi-channel: meaning emerges from the coordinated interplay of hands, facial expressions, head movements, eye gaze, mouthings, and body posture, with non-manual signals often acting as grammatical markers rather than optional visual context. Recognition is also fundamentally temporal: signs unfold over time, and practical systems must model motion, coarticulation, and, in continuous settings, segmentation between sign boundaries. These challenges are compounded by occlusions and viewpoint changes, since hands can occlude each other or the face, and camera placement strongly affects visibility and appearance. From a modeling standpoint, these properties motivate spatio-temporal architectures, multimodal fusion when available (e.g., depth, inertial sensing), and sequence models designed to tolerate real-world variability.

A further linguistic challenge concerns the gap between isolated sign recognition and real-world conversational translation. Isolated SLR assumes that each clip contains a pre-segmented sign with a relatively clear label, whereas natural signing is continuous, context-dependent, and grammatically multimodal. In continuous signing, temporal boundaries are not explicit, and signs are affected by coarticulation: handshape, location, orientation, and motion trajectory may change depending on the preceding and following signs. Non-manual markers add another layer of complexity, because facial expression, eyebrow movement, eye gaze, mouthings, head motion, and body posture can encode negation, questions, emphasis, role shift, temporal structure, and pragmatic intent. These aspects make the transition from isolated medical vocabulary recognition to full patient-doctor dialogue substantially more difficult than a simple extension of clip-level classification. In this dissertation, the main focus is therefore on controlled medical-domain recognition and privacy-preserving sensing, while conversational translation is treated as a longer-term objective requiring explicit modeling of segmentation, context, non-manual markers, and dialogue-level semantics.

In healthcare environments, privacy and acceptability constraints become decisive deployment barriers. Video recordings can capture faces, the surrounding clinical environment, and bystanders, raising concerns even when the intent is strictly assistive. These concerns restrict data collection, complicate governance, and can reduce user acceptance, limiting translation even for technically strong models. A central direction explored in this dissertation is therefore privacy-preserving sensing as an architectural choice rather than a post-hoc constraint. mmWave radar, for example, can capture motion patterns and hand dynamics

while reducing direct identifiability relative to RGB video. This design choice aligns with a broader deployment-oriented viewpoint: changing the sensing modality can unlock practical pathways that would otherwise be blocked by privacy constraints.

Progress in medical SLR is further limited by the scarcity of datasets that reflect realistic clinical communication, including domain-specific vocabulary and multimodal sensing streams. Evaluation protocols must balance comparability with clinical relevance, emphasizing robustness, cross-signer generalization, and error characterization rather than only average accuracy. Community benchmarks and challenges can accelerate progress by standardizing datasets, metrics, and baselines, and by exposing strengths and failure modes under shared conditions. Accordingly, this dissertation includes both dataset contributions and community-level benchmarking results through an organized challenge, providing an empirical perspective on current capabilities and open problems.

Part III addresses these needs through a progression of contributions. It first introduces multimodal datasets and sensor-fusion strategies to support reproducible research in healthcare-oriented SLR. It then develops privacy-preserving recognition using radar-based sensing and investigates semantic grounding through text-aligned radar representations. It also reports insights from a community challenge, emphasizing evaluation practices, robustness, and the role of shared benchmarks in advancing inclusive medical communication.

Chapter 16

Related Work of SLR

Automatic SLR has progressed substantially over the last decade, driven by advances in spatio-temporal representation learning, the availability of larger annotated datasets, and a growing focus on deployment contexts where privacy, robustness, and usability constraints are as important as raw recognition accuracy. Within the scope of this dissertation, the literature can be organized along three complementary axes: sensing technologies, learning paradigms for isolated and continuous recognition, and dataset resources. A recurring theme across these axes is the gap between widely-used visual corpora and the still-limited scale, diversity, and benchmarking maturity of radar-based SLR datasets.

A first dimension concerns sensing technologies. Early SLR systems frequently relied on specialized wearable hardware, such as instrumented gloves, to capture precise hand kinematics [349]. Although accurate, these devices are intrusive, may alter signing behavior, and do not naturally capture non-manual components (e.g., facial expressions, head pose, and body posture) that are structurally essential in sign languages. To overcome these limitations, vision-based approaches have become predominant, leveraging RGB, depth, infrared, and skeleton representations extracted from cameras [196, 55]. Their core advantage is the ability to represent both manual and non-manual channels and thus model the simultaneous, multi-component nature of signing [376]. A wide range of deep architectures has been explored in this setting, including CNNs, RNN/LSTMs, temporal convolutions, graph-based models over pose sequences, and Transformer-based designs for long-range temporal modeling [95, 234, 287, 381, 33]. Multimodal visual pipelines that combine RGB-D, optical flow, and skeletal streams have repeatedly shown improved robustness, particularly when fine hand motion and global body context must be modeled jointly [306, 76, 368]. Despite their effectiveness, camera-centric solutions are often problematic in real settings: in sensitive environments, especially in healthcare, continuous video acquisition and storage raise privacy and governance concerns, and performance can degrade under background clutter, low light,

occlusions, and variable viewpoints [186, 220, 125]. These limitations have motivated radio-frequency (RF) and radar-based sensing as a privacy-preserving alternative. Radar systems can encode motion without capturing identifiable appearance and are robust to illumination changes and many background variations. A broad literature explores RF modalities, including UWB, Doppler and micro-Doppler, continuous-wave radar, FMCW radar, multi-antenna configurations, and mmWave sensing [125, 283, 195, 186, 229, 310, 385, 225, 3, 394]. Recent contributions also consider higher-fidelity representations (e.g., range-Doppler, range-angle, and their temporal dynamics) and direct learning from raw IQ streams, aiming to reduce handcrafted preprocessing assumptions [55, 28, 29]. Overall, radar-based recognition offers an attractive compromise between usability and confidentiality, yet it remains constrained by data scale, vocabulary coverage, and limited public benchmarks.

A second dimension concerns recognition paradigms, commonly distinguished into isolated sign language recognition (ISLR) and continuous sign language recognition (CSLR). ISLR addresses classification of pre-segmented clips and is directly relevant to command-like interactions and many clinical vocabulary settings. Modern ISLR spans video-based pipelines on RGB/RGB-D and skeleton streams, as well as pose-driven or keypoint-driven pipelines that represent signing through 2D/3D trajectories of face, hands, and body. The latter can improve invariance to appearance and lighting and has motivated graph convolution and Transformer-based sequence models [196, 33]. Recent state-of-the-art ISLR increasingly adopts multi-stream fusion, combining global body motion, local hand/face appearance, and skeletal dynamics, confirming that complementary cues can be decisive for fine-grained discrimination [227, 76, 368]. CSLR, in contrast, requires modeling unsegmented signing with temporal alignment, coarticulation, and variable signing rates. A rich line of work combines spatio-temporal feature extraction with alignment objectives such as CTC and weakly supervised learning, while more recent methods employ attention and Transformer-based sequence modeling to improve long-range dependencies and recognition stability [67, 57, 42, 425]. Current CSLR research also explores stronger cross-modality constraints and refined evaluation protocols, including metrics and losses that address frequency imbalance in gloss sequences, to enable fairer comparisons across methods and datasets [179]. These advances remain relevant even when the target application centers on isolated vocabularies, because they clarify how models behave under streaming conditions and under domain shift.

The distinction between isolated and continuous recognition is not only technical, but also linguistic. ISLR benchmarks are useful for evaluating discriminative models under controlled vocabularies, but they simplify several phenomena that are central to natural sign language. In CSLR, temporal segmentation boundaries are usually not provided, signs are modified by coarticulation, and the same manual form can change meaning depending on non-manual

markers and discourse context. Sign language translation introduces an additional level of difficulty because the target is not merely a gloss sequence, but a sentence-level interpretation that must preserve meaning, grammar, and pragmatic intent. This is particularly relevant in clinical communication, where misunderstanding a symptom, a negation, a temporal expression, or an intensity marker can have safety implications. For this reason, progress from medical ISLR toward conversational healthcare translation requires models that jointly encode manual articulators, non-manual components, temporal context, and semantic grounding, together with evaluation protocols that go beyond isolated-label accuracy.

Within both ISLR and CSLR, privacy-aware sensing introduces additional modeling questions. Radar-based recognition is typically closer to motion recognition than to full sign language understanding because non-manual signals are less explicit, motivating architectures that better exploit radar dynamics and fusion strategies that incorporate less identifiable auxiliary sensors. Prior work has also considered domain adaptation and kinematic constraints to address differences between signers and between imitation versus native signing, reinforcing that generalization is a central issue for RF-based SLR [297, 298]. These observations motivate benchmarks that explicitly stress-test signer variability, viewpoint changes, and deployment conditions, rather than evaluating only in controlled, single-signer setups.

A third dimension concerns datasets and resources, with particular emphasis on LIS and radar. Progress in SLR is strongly tied to dataset scale, annotation quality, and coverage of signer diversity. Large RGB-based datasets have enabled data-hungry deep models, including WLASL [196] and MS-ASL [171] for ASL, alongside resources for other sign languages such as SIGNUM [377] and PHOENIX-Weather [108]. Multimodal datasets providing additional streams, including depth and skeleton data acquired with sensors such as Kinect, have supported systematic studies of modality fusion and robustness; representative examples include AUTSL [339] and DGS Kinect datasets [361]. In the Italian context, the creation of LIS corpora and datasets is an active area, yet many resources are private, oriented toward linguistic documentation, or centered on continuous signing rather than clip-level isolated vocabularies used in recognition benchmarks [376, 50, 113, 321]. Methodologies vary substantially (from elicitation and naming tasks to semi-structured interviews and educational recordings), reflecting different scientific goals and leading to heterogeneous annotation practices [45]. While initiatives such as SpreadTheSign provide broad accessibility and value for documentation, they are not designed as standardized machine learning benchmarks and typically lack synchronized multimodal sensor streams [143]. As a result, a gap remains between linguistic corpora and the needs of modern deep models for reproducible evaluation in domain-specific applications such as medical communication.

Compared to camera-based resources, radar-based datasets for SLR are still limited in

vocabulary size and diversity. Many published collections focus on small sets of signs or generic gestures, often with repeated trials by a small number of participants [229, 220, 297]. This scale mismatch restricts the ability to train and compare deep models under realistic conditions, and hinders research on generalization across signers, environments, and sensor placements. Consequently, multiple works argue that radar should be incorporated into larger multimodal datasets to combine privacy-preserving sensing with richer supervision and more meaningful evaluation protocols [125, 297].

This landscape motivates datasets and benchmarks that target clinically relevant vocabularies, support multimodal fusion across visual and non-visual streams, and enable privacy-preserving research directions without relying exclusively on raw video. In this dissertation, the *MultiMeDaLIS* dataset is positioned as a step toward this objective by focusing on medical-domain LIS signs and enabling evaluation across multiple modalities (e.g., RGB, depth, optical flow, skeleton, and radar), together with benchmarking of representative deep models on visual streams and the development of radar-centric pipelines for privacy-aware recognition. This positioning aligns with the broader need, highlighted across the literature, to move from proof-of-concept radar studies on small lexicons to reproducible, scalable SLR systems that can function under real deployment constraints in healthcare.

Chapter 17

Sign Language Recognition for Patient-Doctor Communication: A Multimedia/Multimodal Dataset

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Gaia Caligiore, Concetto Spampinato, Sabina Fontana, Simone Palazzo, and Egidio Ragonese. Sign language recognition for patient-doctor communication: A multimedia/-multimodal dataset. In *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 202–207. IEEE, 2024

17.1 Overview

Sign language translation can substantially improve accessibility and quality of life for Deaf communities, yet it remains a technically demanding problem because meaning is conveyed through multiple coordinated channels, including hand gestures, facial expressions, and upper-body dynamics, all evolving over time. In the clinical setting, this challenge becomes even more consequential: effective patient-doctor communication requires a domain-specific vocabulary and robust sensing under realistic constraints. Within this context, this work introduces *MultiMeDaLIS*, a multimedia and multimodal dataset designed to support LIS recognition and translation for medical communication through sensor fusion.

LIS, officially recognized in Italy in May 2021, is a natural visual-manual language used by a large Deaf population. Like other sign languages, LIS is intrinsically multimodal and multilinear: manual components (handshape, orientation, location, movement, potentially involving one or both hands) co-occur with non-manual components such as facial expressions,

mouth actions, torso motion, and gaze, which jointly contribute to meaning. A further practical challenge is that LIS lacks a standardized writing system, and its oral and visual modality makes documentation and large-scale resource creation difficult. As a result, LIS remains under-documented, with limited publicly available data collections, which directly constrains the development and evaluation of machine learning models for sign language recognition and translation.

To address this gap, the work develops a synchronized, multi-sensor acquisition pipeline that captures complementary views of signing. The dataset combines mm-wave radar, lidar, and RGB/RGB-D cameras to record face, hands, and body information concurrently. This sensor-fusion perspective is motivated by both robustness and deployability: while RGB/RGB-D can capture rich appearance cues, alternative sensors such as radar and lidar can provide motion-related and geometry-related information that remains informative under challenging visual conditions and may support privacy-aware deployment scenarios. The resulting multimodal recordings are intended to enable research on fusion strategies and to provide a basis for recognition models that exploit the full multimodal nature of LIS.

The dataset is focused on a medical semantic domain relevant to patient-staff interactions. The selected vocabulary includes 126 isolated elements, comprising 100 lexical signs plus the 26 LIS alphabet signs, for a total of 25,830 sign instances. Sign selection follows practical standards by including both health-related signs (symptoms, conditions, healthcare concepts) and additional unmarked signs that are useful for composing utterances expected in real clinical exchanges. To improve reliability and consistency of the captured variants, the dataset adopts sign references from the SpreadTheSign dictionary, while the inclusion of medical terminology is informed by a review of Italian and international studies on post-2020 Covid-19 symptom and terminology usage [230]. Overall, *MultiMeDaLIS* aims to provide a structured and reproducible resource for multimodal LIS research, explicitly aligned with the needs of medical communication and with the broader goal of enabling practical translation systems through robust, sensor-aware data acquisition.

17.2 Data acquisition and methodology

17.2.1 Selection of Signs

As already mentioned in Section II, a decision was made to semantically limit the signs included in the dataset to the medical domain. The recorded signs are consequently divided into macro categories: semantically marked and unmarked signs. All signs were recorded in isolation. In identifying the signs to include in the dataset, it was decided to consider other

existing corpora resources for LIS [49, 50, 113, 115] and spoken Italian [217, 272], as well as frequency vocabularies for spoken Italian [80].

To our knowledge, no digital collections of LIS with a specific focus on the medical domain are available. Consequently, in the absence of collections of this kind, we sourced semantically marked terms for the *MultiMeDaLIS* dataset through a review of common COVID-19 symptoms in both Italian and English [313, 4]. Additionally, terms related to both diagnosis and treatment solutions are considered [272]. To deepen our understanding of the experience of Italian deaf people in medical contexts, a set of interviews was conducted respectively with a doctor, a deaf person, and a sign language interpreter.

For the sake of completeness, Table 17.1 summarizes labels of the selected signs, grouped in various categories.

17.2.2 Selection of Tools for Data Collection

Previous experiences in data collection of sign-language data highlight how crucial are both the identification of tools for capturing the *MultiMeDaLIS* dataset and a well defined methodological approach. The visual-manual modality of LIS accentuated the importance of capturing both manual and body components and their positions in the 3D space.

The adopted methodological approach has been grounded in the following key principles to select the recording tools:

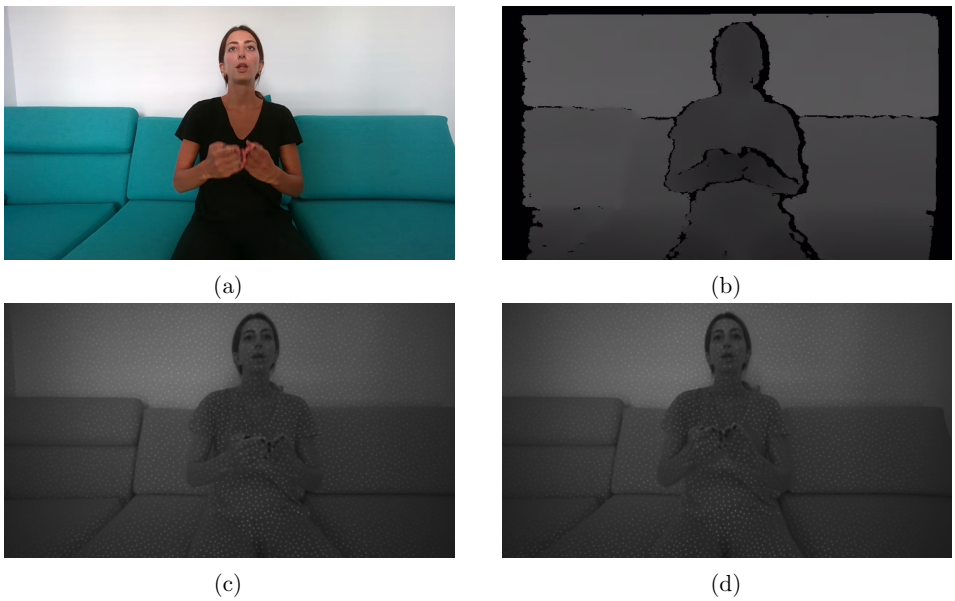
- *portability*: the designed set-up should facilitate data acquisition across diverse settings, enabling signers to easily express themselves in real-world contexts;
- non-invasiveness: the data acquisition process should not interfere with the signer's natural movements or comfort, ensuring that the collected data reflects authentic sign language usage;
- *multimodality*: different sensor technologies provide specific advantages and shortcomings; hence it is critical to integrate multiple acquisition modalities to fully represent the complexity of sign language communication;
- *replicability*: as explained above, the protocol for data collection should be suitable to be employed by users in their everyday life, without introducing data distribution shifts;
- *high-quality data*: the recording equipment must be capable of capturing data with an elevated level of detail and clarity, to facilitate accurate analysis and interpretation of the sign language.

Table 17.1: Labels of the selected signs grouped by category

Categories	**Sign English Labels**		
Body parts	Arm	Back	Chest
	Ear	Eye	Feet
	Forehead	Hand	Head
	Heart	Mouth	Muscle
	Neck	Shoulder	Throat
Medical actions/Conditions	Bad	Cough	Anxiety
	Deaf	Difficulty	Illness
	Infection	Injection	Medicine
	Nausea	Pain	Pale
	Positive	Stomach	Sweat
	Symptoms	Tiredness/Fatigue	Vertigo
	Vomit		
Medical facilities/Professionals	Doctor	Emergency room	Hospital
	Pharmacy		
Time expressions	Day	Morning	Night
	Noon	Time	Tomorrow
	Yesterday		
Health-related terms	To ask	To breathe	To buy
	To call	To communicate	To cure
	To disinfect	To eat	To have a cold
	To have to	To smoke	To stay
	To stop		
Pronouns	Me	You	
	Father	Mother	Son/Daughter
Interrogatives	How	When	Where
	Why		
Negation	No		
	A lot		
Conjunction	If		
	Yes		
Instruments/tools	Mask	Syrup	Thermometer
	Vaccine		
Other	Ambulance	Emergency room	Groin
	Heat	Helicopter	House
	Light	Name	Noise
	Person	Problem	Shoulders

The dataset has been built by collecting synchronized data recorded by means of the following tools:

- Infineon XENSIV 60 GHz RADAR [153];
- Intel Realsense D455 depth camera[154];



(a) (b) (c) (d)

Figure 17.1: Data acquisition: (a) RGB, (b) depth map; (c) infrared stereoscopic with laser projection (left vision), (d) infrared stereoscopic with laser projection (right vision).

- Microsoft Kinect v1 motion camera [235];

- Stereolabs Zed 2 stereo camera [347].

RADAR technology enables capturing the motion and position of the signer's hands and fingers (i.e., manual data), while ensuring users privacy since the visual image is not recorded. RADAR is also able to capture subtle movements, works in absence of physical contact and is not affected by ambient light, allowing for flexibility of recording environments. For these reasons, the use of high-resolution mm wave radar gives the dataset a high degree of novelty. Realsense D455 depth camera is used to record several types of body data, while facial expressions are captured by means of Kinect v1. Finaly, Zed 2 camera ensure high resolution data capture, comprehensively encompassing both manual and body components.

For the sake of completeness, Table 17.2 details each adopted recording tool and the corresponding captured data.

Fig. 17.1 shows the acquired data with different technologies, while the range doppler map from the three radar antennas is depicted in Fig. 17.2. The body and face tracking output are reported in Fig. 17.3.

Table 17.2: Recording tools and corresponding acquired data

Recording Tool	**Captured Data**
Infineon XENSIV 60 GHz RADAR	Various types of manual data at 13 fps, including: • Time-Domain data • Frequency-Domain data • Range Doppler Map data • Range DoA Map data • Moving Target Indication (MTI) collected at the same frame rate
Intel RealSense D455 Depth Camera	Various type of body data, including: • n. 2 infrared (IR) pictures in stereoscopic mode at 848×480 pixels, 30 fps • n. 1 RGB picture at 1280×720 pixels, 30 fps • n. 1 16-bit depth map at 848×480 pixels, 30 fps • n. 68 face tracking points at 30 fps
Microsoft Kinect v1 Motion Camera	Various data types specifically related to facial expression: • n. 1 RGB image at 640×480 pixels, 30 fps • n. 1 16-bit depth map at 640×480 pixels, 30 fps • 87 2D face tracking points at 30 fps
Stereolabs ZED 2 Stereo Camera	Employed to capture the following data types: • n. 2 RGB images in stereoscopic mode at 1920×1080 pixels, 25 fps • n. 2 32-bit depth maps at 1920×1080 pixels, 25 fps • 18 3D body tracking points at 25 fps

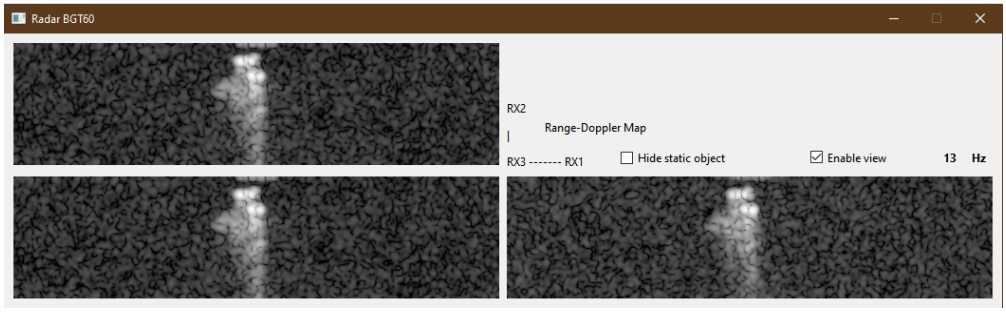


Figure 17.2: Range doppler map from the three radar antennas.



(a) (b)

Figure 17.3: Face (a) and body (b) tracking output.

17.2.3 The Recording Protocol

To ensure the consistency and reliability of the collected data a recording protocol was developed. A total of 200 recording sessions were executed to strengthen the robustness of the dataset. Additionally, five extra sessions were conducted to account for any recordings that were unsuccessful or inadequate, thereby ensuring a complete set of 200 high-quality recordings. Although the protocol is designed to be reproducible by multiple individuals, it has been initially validated and subjected to preliminary analysis using AI models with a single non-native LIF signer who possesses extensive experience in Italian Sign Language.

To facilitate data recording during data collection, a customized software application was designed. This software was uploaded to a desktop computer and a keyboard was repurposed into a two-pedal control device. The pedal board was equipped with two pedals, one for forward navigation (right pedal) and one for backward navigation (left pedal) within the data collection software. This arrangement allowed users to control the data recording process by using their feet, minimizing upper body movement, and reducing the potential for altering the neutral recording position.



(a) (b)

Figure 17.4: Background color variation to differentiate the preparation (yellow) and recording (green) phases of the sign.

During data collection sessions, the signer is presented with one of the 126 Italian labels or letters of the Italian alphabet on the computer screen. Each sign recording begins with the participant in a neutral posture, and they must return to this position after completing each sign. The signer uses the right pedal to indicate the completion of the sign. However, if the participant makes an error (e.g., produces an unclear sign or presses the right pedal prematurely), they can use the left pedal to restart the recording for that specific sign, allowing for re-recording as many times as needed.

The software employs background color variation to differentiate the preparation and recording phases. A yellow background marks the preparation phase, giving participants time to position themselves for the upcoming sign, while a green background marks the recording phase, prompting the participant to begin signing, as shown in Fig. 17.4.

17.2.4 Discussion on Usability and Reliability

We assess the reliability of the proposed dataset through validation sessions with ten native LIS signatories from various regions of Italy, from north to south. This group includes signers of different ages, genders and regional backgrounds. Participants ranged in age from 18 to 54: 40% were 35 to 44, 30% were 25 to 34, 20% were 45 to 54, and 10% were 18 to 24. The group was made up of 6 men and 4 women, all deaf signers, coming from different regions of Italy such as Lombardy, Veneto, Tuscany, Marche, Lazio, Campania, Sicily and two from Puglia, with one participant from Tunisia. Overall, signers provided a positive feedback, confirming the accuracy and cultural relevance of the signs collected, ensuring that the dataset is technically valid and socially and linguistically appropriate. In order to favour usability, in addition to carrying out the next phase of expanding the dataset and consequent publication

of a new artificial intelligence method, the complete dataset will also be released, sharing with the scientific community a valuable resource for further research and development in the field.

17.3 Discussion

The work reported in this paper establishes an initial but concrete step toward medically oriented Sign Language Recognition and translation for patient-doctor communication, where both domain specificity and acquisition realism matter. By collecting a LIS vocabulary of 100 medical signs plus the 26 alphabet signs, with 205 repetitions per sign from a native signer, the study delivers, to the best of our knowledge, one of the first multimodal datasets exceeding 25,000 isolated sign instances explicitly targeting the clinical semantic domain. A key element is the sensor-fusion acquisition strategy: synchronized streams from mm-wave radar, lidar, RGB/RGB-D, and stereoscopic cameras capture complementary aspects of signing across face, hands, and body, enabling future research not only on recognition performance but also on modality redundancy, failure modes, and fusion design under practical constraints.

The preliminary experiments, conducted with established machine learning and deep learning methods on individual acquisition channels (e.g., RGB, depth, and body-tracking joints), provide a useful reference point. While these results are necessarily limited by the single-signer setting and the isolated-sign protocol, they serve an important purpose: they validate the internal consistency of the acquisition pipeline, identify baseline difficulty across sign categories, and offer initial benchmarks for subsequent multimodal models. In this sense, the main value of the present contribution is not a finalized recognition system, but a reproducible data and sensing foundation that enables method development in a clinically relevant scope.

At the same time, several limitations define the immediate research agenda. First, generalization cannot be assessed adequately with one signer, even with many repetitions: signer-specific style, articulation speed, and coarticulation patterns can dominate model behavior in ways that do not reflect real deployment. Extending the protocol to additional native signers is therefore not only a scale-up step but a methodological requirement. The planned replication with ten additional signers is expected to introduce variability that is essential for training and evaluating robust models, as well as for analyzing which modalities best mitigate signer-dependent bias.

Second, the broader translation goal introduces challenges that isolated-sign classification does not fully capture, including continuous signing, segmentation, and context-dependent meaning. The current dataset nonetheless provides a practical starting point by defining

a controlled medical lexicon and by enabling systematic exploration of sensor fusion in a setting where privacy and acceptability are central. In particular, the multimodal design supports a clinically motivated direction: developing models that operate primarily on radar at inference time, while using depth and skeletal/joint signals during training as auxiliary supervision to improve representations and generalization. This teacher-student or privileged-information perspective is appealing because it can reduce reliance on identifiable visual cues in deployment while still exploiting richer modalities in controlled training environments.

Finally, the dataset design directly connects to hospital-specific constraints that often inhibit real-world adoption. Consent management, privacy concerns linked to facial imagery, and governance risks associated with data transfer or non-local processing are not peripheral issues in clinical SLR; they are dominant constraints that shape what can be acquired, stored, and deployed. By incorporating mm-wave radar as a first-class modality and by enabling future pipelines where sensitive visual streams are minimized or avoided at inference, the proposed direction supports a pragmatic pathway toward privacy-aware assistive technologies. In this sense, the work motivates a research trajectory in which performance improvements are coupled with privacy-preserving sensing and deployment feasibility, rather than pursued in isolation from clinical constraints.

Chapter 18

Multisource Approaches to Italian Sign Language (LIS) Recognition: Insights from the MultiMeDaLIS dataset

Source note. The following chapter is based on and adapted from:

- Gaia Caligiore, Raffaele Mineo, Concetto Spampinato, Egidio Ragonese, Simone Palazzo, and Sabina Fontana. Multisource approaches to italian sign language (LIS) recognition: Insights from the multimedalis dataset. In Felice DellOrletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 132–140, Pisa, Italy, December 2024. CEUR Workshop Proceedings

18.1 Overview

As unwritten visual-gestural languages, sign languages require video-based documentation and analysis, which has increasingly pushed automatic recognition research toward multisource acquisition and multimodal modeling. Within computational linguistics, the recognition of sign languages (SLs) as fully-fledged linguistic systems has matured since the 1960s, starting from foundational contributions that established SLs as natural languages rather than mere gestural codes [349]. In the Italian context, the study of signed communication began in the early 1980s through joint efforts of Deaf and hearing researchers, at a time when what is now referred to as LIS was often informally labeled as “mime” or “gesture” by both signers and non-signers [376]. Early publications on LIS [257, 375] and subsequent collaborative work supported a broader shift in perspective and research practices [107]. From

a societal standpoint, accessibility initiatives in broadcast media and public communication have also contributed to LIS visibility, with the COVID-19 period further highlighting the need for equitable access to critical information and the role of LIS in reducing communication barriers [358]. This evolving landscape reinforces a methodological point that is specific to SLs: the absence of a standardized writing system makes high-quality video collections, and their structured annotations, a prerequisite for both linguistic analysis and data-driven recognition.

In this work, we investigate isolated LIS sign recognition in the medical domain using the *MultiMeDaLIS* dataset. We study how combining complementary visual sources can improve recognition robustness by leveraging RGB frames, depth data, optical flow, and skeletal information. Building on these representations, we develop and evaluate two computational approaches that reflect prominent modeling families for sign and action understanding: a Skeleton-Based Graph Convolutional Network (SL-GCN), which exploits the relational structure of body joints over time, and a Spatiotemporal Separable Convolutional Network (SSTCN), which captures temporal dynamics through efficient factorized spatiotemporal processing. While radar data were collected as part of the broader dataset effort, the experiments reported here focus on the aforementioned modalities and do not include radar in the testing phase.

Overall, the study provides evidence that multisource inputs, when structured into suitable representations such as optical flow and skeleton graphs, can improve accuracy and stability for isolated medical LIS sign recognition. From a computational linguistics perspective, these results support the broader view that multimodal datasets and fusion-ready pipelines are a practical direction to improve linguistic accessibility and inclusivity, especially in domain-specific settings where reliable recognition can directly reduce communication barriers for members of the signing community.

18.2 Testing the Dataset

The *MultiMeDaLIS* dataset was designed with the aim of supporting the development of SLR models by enabling the collection and integration of information through various data modalities:

- RGB frames: images extracted from videos.
- Depth data: three-dimensional information for each RGB frame
- Optical flow: to emphasize movement

- Skeletal data: face landmarks and body joints

One of the main components of the dataset are RGB frames, which are images extracted from videos. These frames provide a two-dimensional visual representation of the signs performed by the signer, capturing details such as hand positions and facial expressions. The dataset includes depth data, providing a three-dimensional aspect to the images. allowing for more detailed information on the distance and relative position of elements in the scene. This type of data is particularly useful for understanding the spatial dynamics of signs.

Alongside RGB and depth data, the *MultiMeDaLIS* dataset also contains optical flow information, which describes the movement between consecutive frames. Optical flow is essential for capturing the direction and speed of movements, providing a more detailed understanding of the transitions between various signs. Finally, the dataset includes skeletal data, representing face landmarks and body joints, allowing for precise tracking of joint and body segment positions, facilitating the analysis of signs in terms of joint movements.

Managing this multimodal data is an emerging topic in computational linguistics. By combining different sources of information, it is possible to significantly improve the performance of SLR models. For example, integrating depth data with RGB frames can provide a more complete representation of signs, while adding optical flow and skeletal data can further enrich the analysis of movement's temporal structure. In our view, the *MultiMeDaLIS* dataset provides a solid foundation for exploring these combinations, allowing researchers to develop more effective and accurate solutions for SLR.

18.3 Models and Architectures

In the context of automatic SLR, various approaches and model architectures have been tested to leverage the characteristics of multimodal data in the *MultiMeDaLIS* dataset.

The SL-GCN (Skeleton-Based Graph Convolutional Network) represents a significant innovation in this field. This model generates skeletal data from videos and creates temporal graphs that capture the spatiotemporal relationships between joint movements. Through fine-tuning and the combination of different data streams, SL-GCN has demonstrated high accuracy in sign recognition [74, 263].

Another prominent architecture is the SSTCN (Spatiotemporal Separable Convolutional Network) [165], which excels in feature extraction from videos using HRNet [352]. This approach has shown an accuracy of 96.33%, highlighting its effectiveness in capturing spatial and temporal dynamics of LIS signs.

RGB frames are crucial for the visual representation of signs. The process of splitting videos into frames, cropping, and normalization optimally prepares the data for analysis by

deep learning models. The use of dense optical flow presents significant challenges in sign recognition. Optical flow extraction using the Farneback algorithm [96] led to 56% accuracy, highlighting difficulties in capturing precise details of movements, alongside computational limitations. Depth data encoded with Height, Horizontal disparity, Angle (HHA) represent another crucial resource in the *MultiMeDaLIS* dataset. Applying HHA encoding to depth frames achieved 88% accuracy using the ResNet(2+1)D architecture [364], substantiating importance of three-dimensional information in enhancing understanding and interpretation of signs, offering a more detailed perspective compared to two-dimensional data.

18.4 Training and Evaluation Procedure

For the training of the models, we employed a multi-stream approach that integrates skeletal, RGB, and depth data to improve sign recognition accuracy. The models were trained on an NVIDIA Tesla T4 16GB GPU using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 8. We applied cross-validation to ensure the robustness of the results, splitting the dataset into training (70%) and validation (15%) subsets and data augmentation techniques, such as color jittering, changing the brightness, contrast, saturation and hue, to increase the diversity of the training data and improve generalization.

The loss function adopted for training was categorical cross-entropy, appropriate for multi-class classification tasks. The models were trained for a maximum of 100 epochs, with an early stopping criterion set to terminate training if no improvement in validation loss was observed for 10 consecutive epochs. For evaluation, we used a test set comprising 15% of the dataset, ensuring that the models were tested on unseen data.

18.5 Results

The results demonstrate the model's efficiency in leveraging multi-modal data for improved outcomes. As can be seen in Table 18.1, the SL-GCN multi-stream model achieved the best accuracy, with a Top-1 accuracy of 97.98% and a Top-5 accuracy of 99.94%, surpassing the performance of models using single data streams such as skeletal joints, bones, or motion alone. This demonstrates the advantage of combining multiple streams of information to capture both spatial and temporal dynamics of signs.

Table 18.1: Performance of SL-GCN multi-stream on the test set.

Data	**Accuracy Top-1 (%)**	**Accuracy Top-5 (%)**
Joints	96.24	99.84
Bones	95.82	99.84
Joint Motion	90.37	99.15
Bone Motion	92.69	99.52
Multi-stream	**97.98**	**99.94**

In Table 18.2, datasets trained on the SL-GCN model are compared. Our dataset produced the highest accuracy (97.98%) among the datasets evaluated, outperforming larger datasets like AUTSL (95.45%).

Table 18.2: Comparison of different datasets on SL-GCN model.

Dataset	**Number of signs**	**Accuracy (%)**
MultiMeDaLIS	126	**97.98**
AUTSL	226	95.45
ASLLVD	20	61.04
Alphabet	26	85.19

Table 18.3 presents a comparison of different methods across the entire dataset. The SL-GCN trained on RGB frames achieved the highest accuracy (97.98%), followed by the SSTCN model with 96.33%. The ResNet(2+1)D architecture showed strong performance when applied to RGB frames (97.29%), but struggled when using optical flow data alone, reaching just 56.31% accuracy, suggesting that while optical flow provides valuable information on motion, it lacks the richness of spatial features found in RGB and depth data. The HHA-encoded depth data, when processed with the ResNet(2+1)D model, achieved an accuracy of 88.04%, confirming that depth information is complementary, but not as effective as RGB data in isolation.

Table 18.3: Performance of various methods on the *MultiMeDaLIS* Dataset.

Methods	**Dataset**	**Accuracy (%)**
SS-CGN	RGB	**97.98**
SSTCN	RGB	96.33
ResNet (2+1)D Optical Flow	RGB	56.31
ResNet (2+1)D Frame	RGB	97.29
ResNet (2+1)D Encoding HHA	Depth	88.04

The results highlight importance of combining multiple data modalities, especially RGB and skeletal data, for improving the accuracy and robustness of SLR systems. The performance of the SL-GCN model with multi-stream data shows the model's ability to effectively capture signs, as well as the dataset's value.

18.6 Discussion

This study represents an initial step toward assessing how the *MultiMeDaLIS* dataset can support multisource SLR in a medically grounded semantic domain. By integrating RGB frames, depth, optical flow, and skeletal representations, the dataset enables the evaluation of modeling choices that rely on complementary cues (appearance, motion, and articulated body structure) which are all central to isolated sign recognition. Within this scope, the experiments provide early evidence that *MultiMeDaLIS* is suitable for benchmarking heterogeneous architectures and for highlighting how different input representations impact recognition performance in a medical vocabulary setting.

From a modeling perspective, the results emphasize the value of structured motion representations. The SL-GCN, operating on skeletal data through temporal graphs, benefits from an explicit encoding of spatiotemporal dependencies among joints, which is well aligned with the compositional nature of signed articulations. These findings suggest that the dataset supports graph-based learning in a stable manner, enabling architectures that exploit relational inductive biases to capture signing dynamics. At the same time, the SSTCN experiments indicate that *MultiMeDaLIS* also provides sufficient variability and temporal structure for convolutional spatiotemporal modeling from visual streams, where separable convolutions can efficiently extract motion-driven patterns from RGB-derived representations. Taken together, these complementary outcomes support a pragmatic conclusion: the dataset is not tied to a single modeling family, but can sustain comparative evaluation across substantially different computational paradigms.

Several limitations and open challenges remain. First, the current evaluation focuses on isolated signs, which simplifies temporal segmentation and reduces the ambiguity introduced by coarticulation, turn-taking, and dialogue context. Clinical interactions, however, often involve continuous signing, interruptions, and variable pacing, which can induce failure modes that are not visible in isolated recognition. Second, although multisource visual modalities improve robustness, they can be constrained in real settings by occlusions, viewpoint restrictions, and privacy requirements. These constraints motivate the broader trajectory of the dataset effort toward alternative sensing and privacy-aware pipelines.

In this direction, the preliminary progress on adapting the models to radar is a particularly relevant step, because radar can reduce reliance on identifiable RGB video while preserving motion information that is useful for recognition. The intended strategy (extracting a radar processing module that can operate independently at inference time) highlights a deployment-driven design choice: removing RGB from the inference pathway can enable resource-constrained, edge-oriented operation, while still allowing richer multimodal supervision during development and training. This separation between training-time multimodality and inference-time deployability is coherent with the broader thesis principle of improving feasibility by redesigning the input channel rather than only increasing model complexity.

Finally, generalization across signers remains a critical requirement for clinical applicability. The planned expansion of the dataset with additional Deaf signers (following the same acquisition protocol) directly addresses signer variability and is necessary to evaluate robustness under realistic inter-subject differences in articulation, speed, and style. In parallel, future work should consider protocol extensions that stress-test deployment factors, including multiple environments, sensor placements, and interaction scenarios, as these choices may influence both performance and acceptability. Overall, the results support the feasibility of multisource SLR in the medical domain and motivate a next phase focused on signer-diverse acquisition, radar-centered inference, and edge-ready system design for emergency or low-resource contexts where access to professional interpreting may be limited.

Chapter 19

Radar-Based Imaging for Sign Language Recognition in Medical Communication

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Gaia Caligiore, Federica Proietto Salanitri, Isaak Kavasidis, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Radar-based imaging for sign language recognition in medical communication. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2025

19.1 Overview

Sign languages are complete visual-gestural systems used by deaf communities worldwide. In Italy, the LIS received formal recognition in May 2021, consolidating decades of linguistic and socio-semiotic research on its structure and usage [36, 376]. In healthcare, however, the availability of LIS as a communication channel remains uneven: misunderstandings and delays can occur when Deaf patients cannot interact effectively with medical staff, and professional interpreters, the ideal solution, may be unavailable in emergency or resource-limited contexts. This gap has motivated increasing interest in automatic SLR as an assistive technology for patient-doctor communication [106]. Yet clinical environments impose constraints that are less prominent in generic SLR settings, with privacy and governance often limiting the use of camera-based pipelines, and with uncontrolled conditions (lighting, clutter, backgrounds) challenging their robustness [186, 220].

Against this backdrop, RF sensing, and in particular millimeter-wave RADAR, offers a privacy-preserving alternative. Unlike RGB acquisition, RADAR captures motion and velocity patterns while obscuring fine-grained visual details that can compromise identity, making it attractive for regulated settings such as hospitals. RADAR sensing is also less sensitive to illumination and background variability, which can be advantageous in dynamic clinical spaces [125, 195]. Nevertheless, existing RADAR-based SLR studies and datasets are often limited in lexicon size and subject diversity, frequently focusing on generic gestures or on ASL-centric settings, which reduces the ability to train and validate deep models that generalize to real medical vocabularies and varied signing styles [241, 41, 40]. This motivates the need for larger, domain-specific resources and for modeling pipelines that explicitly handle temporal variability and deployment constraints.

In this work, we introduce a privacy-preserving framework for LIS recognition in medical contexts, centered on 60 GHz mm-wave RADAR as the primary sensing modality. The task targets 126 isolated LIS elements relevant to patient-doctor interaction (100 lexical medical terms plus the 26 letters of the LIS alphabet) collected under a structured acquisition protocol. While the system is radar-centric at inference, the data collection is designed to be multimodal: synchronized RADAR recordings are aligned with depth/RGB sensors to support analysis, baselines, and future extensions, while maintaining privacy constraints by restricting the core recognition pathway to RADAR-derived motion imaging.

Methodologically, the proposed pipeline follows a two-stage design aimed at separating compact motion representation from temporal decision making. First, a residual autoencoder encodes Range Doppler Maps (RDM) together with MTI channels into low-dimensional embeddings, exploiting skip and residual connections to preserve informative structure while keeping representations efficient. Second, a Transformer-based classifier models temporal dependencies across sequences of varying sign duration, enabling recognition under realistic variability in signing speed and execution. This decomposition supports a deployable inference pathway: the representation stage compresses high-rate radar imaging into compact tokens, while the temporal model captures dynamics without assuming fixed-length sequences.

Experiments conducted on the proposed dataset show that the radar-centric approach can reach high recognition performance in a privacy-constrained setting. In cross-validation, the best configuration achieves 93.6% accuracy, 87.9% sensitivity, 99.3% specificity, and an 87.7% F1 score, outperforming existing approaches including an RGB-based baseline. Beyond sign recognition, these results support a broader message: radar imaging can act as a practical modality for privacy-compliant motion analysis in healthcare, with potential applicability to other movement-centric clinical problems (e.g., neurological or motor disorders) where identity exposure should be minimized. Code is available

at https://github.com/IngRaffaeleMineo/SignRadarClassification_MICCAI2025, and the dataset will be released publicly.

19.2 Method

19.2.1 Dataset Description

LIS encompasses both manual (e.g., handshape, orientation, movement) and non-manual (facial expressions, torso, mouth actions) components [36, 376]. To address patient-doctor interactions, a dataset was compiled featuring 126 LIS items (100 health-related signs plus 26 letters of the alphabet), drawing on previous LIS corpora guidelines, frequency vocabularies for spoken Italian [50], and incorporating medical and deaf-community insights [241, 41]. Following the principles of portability and non-invasiveness outlined in [40], multiple data modalities were simultaneously collected: 13-fps three-antennas RADAR time-domain log-scale data (with an Infineon XENSIV BGT60TR13C 60 GHz sensor); 720p 30-fps RGB-D videos with face tracking points (with an Intel Realsense D455 camera) and additional facial-expression data (through a Microsoft Kinect v1); and 25-fps 1080p depth images/points clouds (using a Stereolabs Zed 2 camera). A single subject performed 205 repetitions per sign, totaling 25,830 sign instances[241, 41].

Tab. 19.1 shows a comparative analysis of our dataset, with respect to existing datasets for SLR. Compared to prior RADAR data collections, our dataset encompasses both a larger number of symbols and repetitions, yielding a significantly larger dataset and a generally more complex classification task. Some RGB and RGB-D datasets (in particular, SpreadTheSign [143] and Li et al. [196]) feature a comparable or larger number of words, but a significantly lower number of repetitions per word, which hinders their suitability for supervised learning. Additionally, in practical applications they are subject to the exposure of privacy-sensitive information.

19.2.2 Data Preprocessing

Our 60 GHz RADAR sensor comprises one transmit antenna and three receive antennas (RX), each recording time-domain signals in logarithmic scale. The radar is configured with a 1 MHz sampling rate, a transmit power level of 31, and an IF gain of 40 dB. Under these settings, the radar achieves a range resolution of 0.0312 m, a maximum range of about 1.60 m, a maximum speed of 4.11 m/s, and a speed resolution of 0.0321 m/s. Each frame is acquired with a repetition time of 0.077 s and a center frequency of 60.50 GHz, ensuring fine-grained

Table 19.1: Summary of existing datasets, in terms of data modality, number of signs, number of subjects and number of repetitions per sign. Notation: for RADAR-only data modality, we report the sensor frequency; for "signs", we report by default the number of sign-language signs, as well as letters ("L") or non-sign-language gestures ("G"). * Plus RGB-D and face tracking data.

Dataset	**Modality**	**# signs**	**# subj.**	**# repet.**
SpreadTheSign[143]	RGB	281,672	Unspecified	1
Li et al.,[196]	RGB	2,000	3	119
Sincan et al.[339]	RGB-D	226	4	43
Ravi et al.[306]	RGB-D	200	10	10
Jing et sl.[168]	RGB-D	100	42	15
Hassan et al.[136]	RGB-D	10	2	22
Li et al.[195]	8.5 GHz	10	10	10
Lu et al.[220]	24 GHz	5	1	220
McCleary et al.[229]	24 GHz	4	5	250
Park et al.[283]	33 GHz	5L + 6G	10	82
Gurbuz et al.[125]	77 GHz	20	3	3
Wang et al.[385]	77 GHz	6G	1	200
Ours[241]	60 GHz*	100 + 26L	1	205

motion capture suitable for sign language gestures. Radar data were preprocessed through range and Doppler FFTs with zero-padding, windowing (Blackman-Harris and Chebyshev for range and Doppler, respectively), mean removal, and a spectral threshold of approximately $-90\,\mathrm{dB}$ (adjustable between $-75\,\mathrm{dB}$ and $-120\,\mathrm{dB}$). An optional MTI filter further reduces static background signals by zeroing out bins below a certain velocity threshold. The resulting maps (both RDM and RDM-MTI) have a resolution of 128×1024.

Signal synchronization was carried out across all sensors (RADAR, RGB-D, depth/point clouds, face features) to ensure time-aligned frames across the 25,830 sign instances, each typically spanning 1-4 seconds. In total, over 6 million radar maps (between RDM/RDM-MTI) are included, providing refined high-resolution range and velocity features for subsequent autoencoder and transformer processing. Please note that our approach only employs RADAR time data from the collected dataset: additional modalities are employed for comparison with methods from the state of the art, or kept for future studies.

19.2.3 Model Architecture

We address the sign language recognition problem as a classification task. Given the nature of RADAR data, we deal with video-like sequences, where each frame represents an RDM or RDM-MTI map, thus requiring the model to capture both spatial and temporal dynamics for

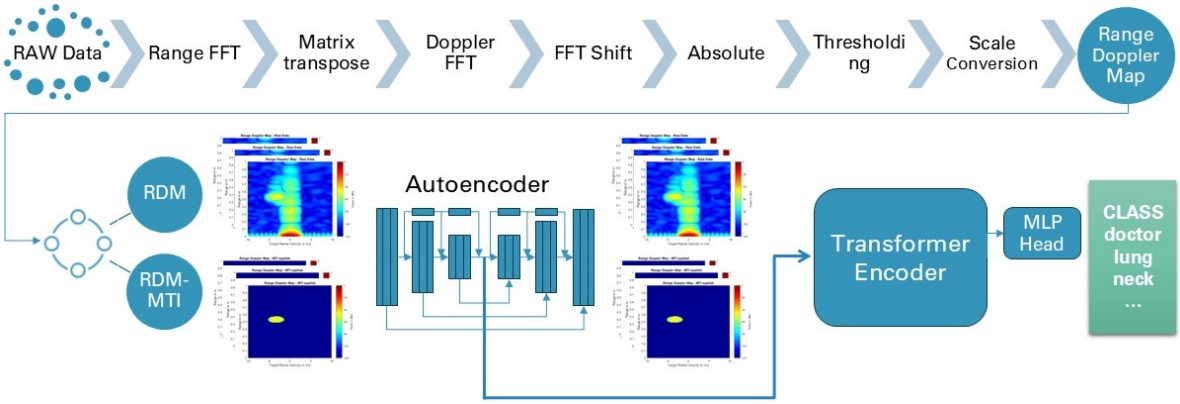


Figure 19.1: Overview of the end-to-end architecture of the proposed method.

effective classification.

Addressing this challenge directly with a single deep network can lead to excessive model complexity and computational costs, particularly given the high dimensionality and sequential nature of radar data. To mitigate these issues, our approach decouples the problem into two distinct stages: frame compression through an autoencoder and sequence classification with a transformer network.

Autoencoder for RADAR feature extraction.

The proposed autoencoder network is designed to learn a compact and meaningful representation of radar data for sign language recognition. The fundamental objective of this autoencoder is to reconstruct the original input RADAR map, either RDM or RDM-MTI, while distilling its essential features into a low-dimensional embedding, which is used later as input to the transformer classifier. Due to the significant differences in nature between RDM/RDM-MTI maps and natural images, we do not employ pretrained models; we thus design a custom CNN architecture[1], with symmetric encoder and decoder, each featuring 9 convolutional layers (grouped into three blocks, for each feature map resolution, before downsampling). We add residual connnections between the input and output of each block, to ease gradient flow. The bottleneck of the model yields a 256-dimensional compressed representation of the input signal.

Transformer classifier

Following the autoencoder pretraining, the classification stage of the proposed approach is based on a transformer network, which aims to effectively classify the sign language represen-

[1]Please refer to the public source code for model details.

tations captured in the 256-dimensional embeddings. The objective of this transformer-based classifier is to leverage its powerful sequence modeling capabilities to capture patterns in the RADAR signal, corresponding to the dynamic gestures of sign language. The transformer architecture is particularly suited for this task due to its inherent ability to model long-range dependencies and capture the sequential nature of sign language gestures. Moreover, it naturally handles variable-length sequences, which is necessary in our scenario due to the different lengths of the recorded sign, ranging from 13 to 66 RDM/RDM-MTI frames.

Our approach utilizes a custom transformer design that processes the output embeddings from the autoencoder. The autoencoder 1024-dimensional embeddings are first projected on a 64-dimensional space; in order to support the combination of multiple RADAR modalities with the same model, whenever multiple modalities are employed we fuse the autoencoder embeddings via convolutional layers by treating them as channels, before feeding them to the initial projection layer. We then augment the sequence with a learnable class token and positional embeddings to maintain temporal context, before being fed to the transformer. The Transformer architecture employs six transformer layers, with standard multi-head self-attention and eight attention heads per layer. At the output layer of the transformer, we feed the resulting class token embedding into a linear classification layer, estimating class logits for the dataset's 126 classes and minimizing a standard cross-entropy loss.

19.2.4 Training procedure

We partition the dataset into 60% training, 20% validation, and 20% test sets. No data augmentation is carried out. As mentioned above, we preliminarily train the autoencoder model, in order to learn a good representation of RADAR signals at a single time step, to simplify the following sequence classification task. We train the model for 15 epochs with a batch size of 16, adopting the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of $5 \cdot 10^{-5}$.

The autoencoder is then frozen and used only to extract the embeddings from the dataset samples. We train the transformer for 700 epochs, with a batch size of 256 and a learning rate of 10^{-4}, with an AdamW optimizer as above. Gradient clipping is enforced at a maximum norm of 5, and a weight decay of $5 \cdot 10^{-6}$ is applied to mitigate overfitting. All experiments are performed in `float16` mixed precision using 5×NVIDIA A100 40GB and 4×NVIDIA A6000 Ada 48GB. This two-stage process alleviates GPU memory constraints by avoiding full end-to-end training on raw data for every epoch.

The above hyperparameters are chosen based on the model's performance on the validation set.

19.3 Experimental Results

In this section, we present a thorough performance assessment of our approach, comparing our results to state-of-the-art methods and justifying architectural choices described above. We report results in terms of classification accuracy, sensitivity, specificity, and F1-score.

Table 19.2 reports the comparison, on our dataset, of the proposed approach with state-of-the-art solutions, drawn from both RGB/RGB-D approaches and radar-based methods. Where the source code is not provided by the authors, we wrote our own implementations, adhering to the reference description at the best of our possibilities. Model selection is performed based on validation performance, using default hyperparameters. For our approach, we report results corresponding to different input modalities (single/multi-antenna, RDM only, RDM-MTI only and the combination of RDM and RDM-MTI).

Table 19.2: Comparison with state-of-the-art methods in sign language recognition. The "Modality" column indicates the primary sensor modality employed by the method.

Method	**Modality**	**Acc.**	**Sens.**	**Spec.**	**F1-score**
De Coster et al. [76]	RGB	88.4 ± 3.7	78.6 ± 3.4	98.2 ± 3.5	79.7 ± 3.9
Vahdani [368]	RGB-D	84.1 ± 2.9	83.1 ± 4.1	85.1 ± 2.1	84.2 ± 2.2
ResNet(2+1)d	RGB-D	74.6 ± 3.9	66.4 ± 3.1	82.8 ± 3.0	59.9 ± 3.3
Jhaung et al. [161]	RADAR	71.9 ± 3.3	61.9 ± 2.3	81.9 ± 3.4	69.2 ± 2.9
Debnath et al. [81]	RADAR	79.3 ± 3.5	83.3 ± 3.3	75.3 ± 3.1	79.3 ± 3.9
Arab et al. [12]	RADAR	81.0 ± 3.4	65.9 ± 3.1	96.1 ± 2.8	29.6 ± 2.9
Ours	RDM	88.3 ± 1.1	79.9 ± 0.8	96.7 ± 0.9	86.6 ± 1.2
	3×RDM	91.7 ± 0.6	86.9 ± 0.6	96.5 ± 0.7	87.5 ± 0.5
	MTI	84.9 ± 0.9	73.9 ± 0.7	95.9 ± 0.8	80.0 ± 0.8
	3×MTI	86.1 ± 0.6	77.6 ± 0.5	94.6 ± 0.5	83.3 ± 0.7
	RDM + MTI	91.4 ± 0.6	84.1 ± 0.5	98.7 ± 0.5	86.9 ± 0.5
	3×RDM + 3×MTI	93.6 ± 0.5	87.9 ± 0.5	99.3 ± 0.6	87.7 ± 0.5

Our approach achieves the best performance when employing three antenna streams and both RDM and RDM-MTI channels, reflecting the advantage of integrating multiple radar perspectives and emphasizing moving-target information. Notably, even though radar systems typically exhibit lower performance compared to RGB-based systems in prior work, our solution achieves comparable, if not better results on several metrics, demonstrating the viability of privacy-preserving, radar-centric solutions for sign language recognition.

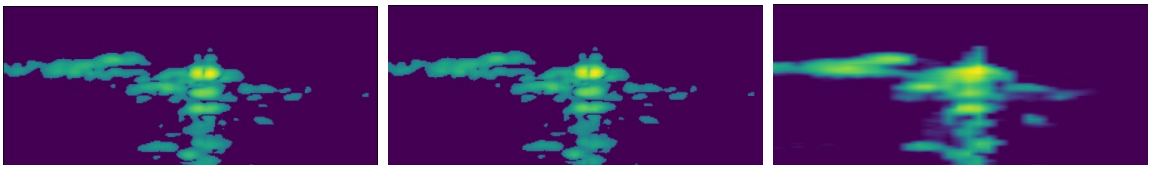
19.4 Ablation Study

19.4.1 Ablation on Autoencoder Reconstruction

We first study how (i) residual/skip connections and (ii) training-set size affect the reconstruction quality of the autoencoder, as reconstruction fidelity is a direct proxy for the quality and stability of the latent embeddings later used for classification. Table 19.3 reports the MSE after 1, 5, and 15 epochs when training with vs. without skip connections, and using either 20% or 100% of the available training data.

Table 19.3: Autoencoder reconstruction error under different conditions. Skip connections consistently reduce MSE, with larger gains as training progresses and as more data are used.

Skip connection	**Dataset (%)**	**MSE Epoch 1**	**MSE Epoch 5**	**MSE Epoch 15**
Yes	20	5×10^{-2}	8×10^{-3}	4×10^{-4}
Yes	100	4×10^{-2}	2×10^{-5}	9×10^{-6}
No	20	8×10^{-1}	3×10^{-1}	8×10^{-2}
No	100	2×10^{-1}	6×10^{-2}	7×10^{-3}



(a) Original RDM (b) Recon. with skip conn. (c) Recon. w/o skip conn.

Figure 19.2: Example reconstructions after 10 epochs. Skip connections preserve fine motion-energy structures that are attenuated when relying on the bottleneck alone

Two trends are evident. First, skip connections yield substantially lower MSE across all regimes, with the gap widening over training, indicating faster and more stable convergence. Second, using the full dataset improves reconstruction quality, especially when skip connections are enabled, suggesting that the residual pathways benefit from richer coverage of the radar data distribution. Overall, higher-fidelity reconstructions translate into cleaner latent embeddings, which empirically improve downstream classification (see the following ablations).

19.4.2 Observations on Skip Connections and Dropout

We observed that removing skip connections increases the number of epochs required to reach a comparable reconstruction quality. Intuitively, in the early stages of training the decoder

can only rely on a still immature bottleneck representation, which slows down convergence. Residual connections alleviate this by providing higher-frequency spatial detail directly to the decoder, allowing the encoder-decoder pair to co-adapt more smoothly from the start.

However, skip connections introduce a well-known risk: if the residual pathway becomes dominant, the network can partially bypass the bottleneck, reducing the information content of the latent code. To prevent this shortcut, we add dropout on the skip pathways. This forces the model, with non-zero probability, to reconstruct using the bottleneck alone, encouraging the latent space to retain task-relevant structure. In practice, this yields a better trade-off between reconstruction fidelity and bottleneck usefulness, and leads to more stable embeddings for the Transformer classifier.

19.4.3 Ablation on Different Backbone Configurations

We further vary the dropout level while keeping skip connections, and evaluate its effect on the final sign classification. Table 19.4 shows that a high (but not total) skip dropout provides the best performance.

Table 19.4: Impact of skip-dropout in the autoencoder backbone, with skip connections always enabled. A high dropout level of 0.8 yields the best overall balance and the highest classification performance.

Dropout level	**Acc.**	**Sens.**	**Spec.**	**F1-score**
0.0	85.5	77.2	93.8	81.3
0.5	88.7	81.8	95.6	81.4
0.8 (Ours)	**93.6**	**87.9**	**99.3**	**87.7**
1.0	91.4	88.9	93.9	85.0

The trend suggests that removing dropout (0) encourages over-reliance on skip pathways and produces less informative bottleneck embeddings. Conversely, fully dropping skip connections (1) removes the reconstruction advantage of residual learning and reduces accuracy and F1. The intermediate regime (0.8) preserves the optimization benefits of residual connections while still forcing the bottleneck to carry discriminative information, which is consistent with the best downstream performance.

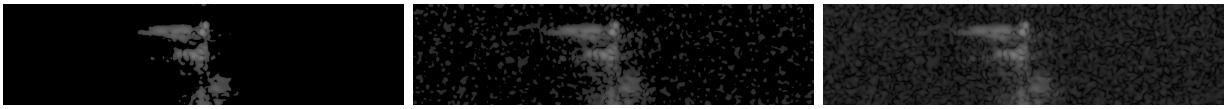
19.4.4 Ablation on RDM Spectrum Threshold

During preprocessing, we apply an RDM spectrum threshold to suppress low-intensity components. This operation trades off noise rejection against the risk of discarding weak but

informative motion patterns. Table 19.5 reports the effect of three thresholds, and Fig. 19.3 provides qualitative examples.

Table 19.5: Effect of different RDM spectrum thresholds. The −90 dB setting provides the best trade-off between suppressing background noise and preserving discriminative motion energy.

RDM spectrum threshold	**Acc.**	**Sens.**	**Spec.**	**F1-score**
−75 dB	92.8	86.6	99.0	87.5
−90 dB (Ours)	**93.6**	**87.9**	**99.3**	**87.7**
−120 dB	90.2	82.3	98.1	84.1



(a) RDM at -75 dB (b) RDM at -90 dB (c) RDM at -120 dB

Figure 19.3: Qualitative comparison of RDMs obtained with different spectrum thresholds

A relatively high threshold (-75 dB) retains more background components, which can slightly increase spurious energy and reduce sensitivity. Conversely, an aggressive threshold (-120 dB) removes weak motion traces and degrades recognition, suggesting that some discriminative cues lie in low-amplitude regions. The intermediate choice (-90 dB) best preserves informative motion structure while suppressing noise, and therefore yields the strongest overall performance.

19.4.5 Ablation on Backbone Type and Pretraining

We finally assess the effectiveness of the proposed autoencoding feature extractor along two axes. First, we compare our residual autoencoder backbone against standard CNN alternatives (AlexNet and ResNet-18). We avoid larger models because RDM images exhibit relatively structured patterns and limited appearance variability, where excessive capacity tends to increase overfitting without clear representational benefits. Second, we evaluate whether fine-tuning the feature extractor jointly with the Transformer classifier improves performance, compared to the default configuration where the backbone is frozen during Transformer training.

Results in Tab. 19.6 show two consistent findings. (i) Freezing the backbone yields better performance across all architectures, indicating that joint fine-tuning increases overfitting (also reflected by the training/validation loss divergence). (ii) Our residual autoencoder

provides the strongest features among the tested backbones, likely because it is explicitly trained to reconstruct radar-specific structure and therefore learns embeddings aligned with the statistics of RDM/MTI inputs, whereas generic CNN backbones are not optimized for this reconstruction-driven inductive bias.

Table 19.6: Comparison of backbone architectures and training strategies (mean ± std). Freezing consistently improves generalization, and our residual autoencoder achieves the best overall performance

Backbone	**Acc.**	**Sens.**	**Spec.**	**F1-score**
AlexNet	61.4 ± 1.6	43.9 ± 1.9	78.9 ± 1.6	54.1 ± 1.3
↪ Frozen	73.9 ± 1.7	68.8 ± 2.0	79.0 ± 2.1	66.6 ± 1.8
ResNet-18	84.3 ± 0.4	76.2 ± 0.4	92.4 ± 0.6	78.2 ± 0.6
↪ Frozen	88.1 ± 1.2	81.7 ± 0.6	94.5 ± 0.9	79.9 ± 0.5
Ours	91.9 ± 0.6	86.1 ± 0.5	97.4 ± 0.6	85.5 ± 0.7
↪ **Frozen**	93.6 ± 0.5	87.9 ± 0.5	99.3 ± 0.6	87.7 ± 0.5

19.5 Discussion

This work supports a clear conclusion: privacy-preserving SLR in healthcare can be pursued effectively without relying on identifiable video streams. By centering inference on 60 GHz RADAR motion imaging and validating on a medical-domain LIS dataset of 126 signed items (100 medically relevant lexical signs plus 26 letters), we show that non-visual sensing can achieve strong recognition performance while aligning naturally with clinical privacy constraints. The proposed two-stage design, a convolutional (residual) autoencoder for compact RDM/MTI representation followed by a Transformer for temporal modeling, is not only competitive in accuracy, but also practically meaningful as an architectural decomposition: representation learning compresses high-rate radar observations into stable embeddings, while the temporal classifier handles variable sign duration and dynamics.

At the same time, several limitations and open directions emerge from the scope of the current study. First, a radar-centric pipeline can struggle with classes that are manually similar and differ primarily through subtle non-manual cues. In medical vocabularies, this case is common: signs referring to related anatomical regions or conditions may share hand configuration and gross motion but differ in mouth actions or facial components. This motivates a natural extension toward incorporating face-joint and lip-joint tracking as an auxiliary signal to improve disambiguation. Importantly, such an extension can be aligned with the same privacy rationale if it relies on landmark-level representations rather than RGB

storage, thereby retaining a low-identifiability input while adding discriminative information for difficult pairs.

Second, while the presented results demonstrate feasibility under controlled data collection and cross-validation, clinical translation requires robustness evidence under broader variability: different signers, sensor placements, room layouts, multipath conditions, and interaction styles typical of hospitals. In this sense, scaling the cohort and expanding acquisition scenarios are not incremental dataset improvements but necessary steps to quantify generalization and to characterize failure modes that matter in deployment (e.g., occlusions, distance changes, and signer-specific dynamics). From a modeling perspective, this also motivates investigating representation learning objectives that better exploit radar-specific dynamics and improve invariance to environmental changes.

Finally, a privacy-preserving SLR system intended for healthcare must be evaluated as an interaction component rather than as an offline classifier. Real-time operation introduces constraints on latency, temporal stability, and user experience: predictions should be streamable, resistant to jitter, and paired with uncertainty-aware behavior that can request clarification when evidence is insufficient. The two-stage design adopted here provides a plausible path toward such deployment, but a full translation to clinical settings will require end-to-end validation with user-centered protocols, including Deaf signers, clinicians, and interpreters, to define acceptable failure modes and appropriate interaction policies.

Overall, these considerations reinforce the main message of the paper: RADAR-based, privacy-preserving sensing is a viable foundation for medical-domain SLR, and future progress is likely to depend on (i) extending data diversity across signers and environments, (ii) integrating low-identifiability non-manual cues such as facial/labial landmarks for fine-grained disambiguation, and (iii) moving toward streaming, uncertainty-aware interaction for real clinical workflows.

19.6 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 775 hours of computation was performed on 4 NVIDIA A100 GPUs (thermal design power of 300 W) and four NVIDIA A6000 GPUs (thermal design power of 300 W), with total emissions estimated[2] to be 91.8 kg CO_2eq.

[2]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 20

Text-Aligned Radar-Based Sign Language Recognition for Healthcare Communication

Source note. The following chapter is based on and adapted from:

- Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4894–4902, 2025

20.1 Overview

SLR aims to bridge the communication gap between Deaf or hard-of-hearing individuals and non-signers through automated interpretation of signs. In recent years, the field has made substantial progress, particularly for *isolated sign recognition*, where each sign instance is processed independently. Early systems relied on handcrafted descriptors and shallow classifiers and showed limited ability to generalize across signers and vocabularies [287, 381]. With the advent of deep learning, end-to-end models have become the dominant paradigm, learning spatiotemporal representations directly from data and leveraging RGB and depth video to capture both manual and non-manual components of signing [196, 33].

Despite their effectiveness, camera-based solutions can be problematic in privacy-sensitive environments such as hospitals and rehabilitation centers. Continuous video monitoring may raise ethical and legal concerns, and performance can be fragile under suboptimal lighting

conditions, cluttered backgrounds, and partial occlusions [186, 220]. These limitations motivate alternative sensing modalities that can preserve privacy by design while remaining robust in real-world settings. Radar-based sensing has recently emerged as a promising direction: frequency-modulated continuous wave (FMCW) radar can capture motion dynamics without exposing identifiable visual information [125, 195], and is inherently resilient to lighting variability and many occlusion scenarios.

Prior radar-based sign recognition studies have demonstrated that dynamic movements can be captured through micro-Doppler signatures and other radar-derived representations [143, 55, 310]. However, much of the existing literature focuses on American Sign Language (ASL) and typically targets small and constrained vocabularies [3, 297]. This narrow scope limits transfer to other sign languages, which can differ substantially in lexicon, grammar, and articulation. In addition, current radar-centric pipelines often lack mechanisms to exploit semantic priors from natural language, missing an opportunity to guide representation learning through textual supervision and to improve discriminability via language-level structure.

To address these gaps, we introduce **TRACE** (*Text-Radar Alignment via Contrastive Embedding*), a radar-based framework for isolated SLR that combines privacy-preserving sensing with language-grounded learning. TRACE extracts spatiotemporal embeddings from sequences of Range-Doppler Maps (RDMs) using a 3D convolutional neural network. During training, it employs **prompt-conditioned contrastive learning** to align radar embeddings with class representations produced by a **frozen text encoder** augmented with **learnable soft prompts** [423]. Specifically, each class label is converted into a textual template whose prompt tokens are optimized jointly with the radar encoder, yielding task-adapted class embeddings while keeping the language model fixed. The training objective encourages matching radar-text pairs to be close in the shared embedding space and non-matching pairs to be separated, thereby guiding the CNN toward semantically meaningful features. At inference time, TRACE assigns to each radar sequence the label whose textual embedding is most similar to the radar embedding, enabling a simple and effective nearest-embedding decision rule.

We evaluate TRACE on the *MultiMeDaLIS* dataset [241], which contains FMCW radar recordings of isolated LIS signs curated for clinical and assistive communication contexts. Experimental results show that TRACE outperforms state-of-the-art radar-based baselines under standard classification metrics. Ablation analyses further indicate that (i) joint multimodal contrastive alignment improves embedding quality by injecting semantic supervision into radar representation learning, and (ii) soft prompts produce more suitable textual class embeddings than fixed prompts, improving generalization and feature discriminability.

The main contributions are threefold: (1) TRACE, a privacy-aware architecture for

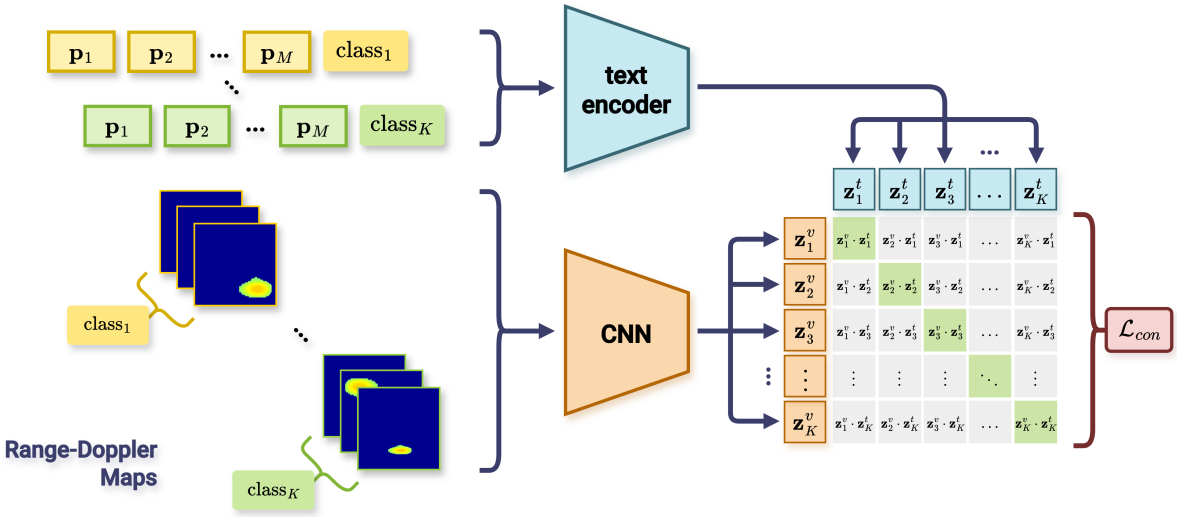


Figure 20.1: Overview of the *TRACE* method. Radar-based RDM sequences are encoded into visual embeddings $\mathbf{z}_i^v$ using a 3D CNN. Each class label is represented textually and prepended with learnable prompts before being processed by a frozen text encoder, yielding text embeddings $\mathbf{z}_k^t$. A contrastive loss $\mathcal{L}_{\text{con}}$ encourages alignment between matching visual–text pairs in the shared embedding space.

radar-based isolated SLR that aligns radar sequences with textual class descriptions through contrastive learning, (2) prompt-conditioned textual embeddings obtained via a frozen language model with learnable soft prompts, enabling adaptable and semantically grounded classification, and (3) a comprehensive evaluation on a radar-acquired LIS dataset demonstrating improved performance and robustness over prior approaches.

20.2 Method

To better capture domain-specific patterns in Italian Sign Language, the proposed method leverages the CLIP text encoder with learnable prompts and employs a 3D convolutional network applied to radar-derived range-Doppler maps as its visual encoder. Figure 20.1 illustrates an overview of the proposed method.

20.2.1 Problem formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ be a dataset, where $\mathbf{x}_i \in \mathcal{X}$ is an input sample and $\mathbf{y}_i \in \mathcal{Y}$ is its corresponding class label. Each input sample $\mathbf{x}_i$ is a sequence of frames $\{\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \ldots, \mathbf{f}_{i,T_i}\}$, where T_i is the temporal length of the video. The goal of our work is to develop a model that maps any given input sample $\mathbf{x}_i \in \mathcal{X}$ to its corresponding class label $\mathbf{y}_i \in \mathcal{Y}$.

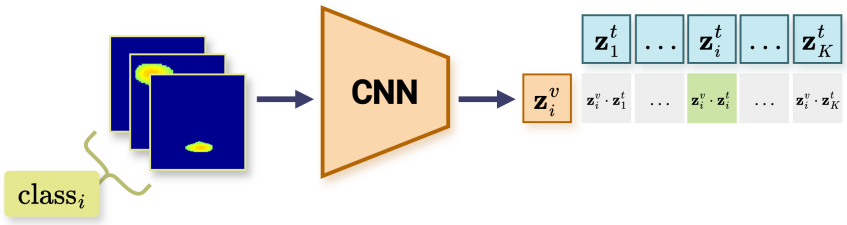


Figure 20.2: Overview of the inference phase. The CNN encodes the RDM sequence into a visual embedding $\mathbf{z}_i^v$, which is compared with all text embeddings $\mathbf{z}_k^t$ corresponding to class labels. The predicted class is the one with the highest cosine similarity between $\mathbf{z}_i^v$ and $\mathbf{z}_k^t$.

20.2.2 Text encoding

Each class label $\mathbf{y}_i \in \mathcal{Y}$ is a short textual string, corresponding to either a medical term or a letter of the Italian alphabet in LIS.

Let $\mathbb{V}$ be a fixed vocabulary; each string is tokenized into $\mathbb{V}^N$, a sequence of words from $\mathbb{V}$. To simplify notation, we treat N as variable-length, leveraging the length-agnostic nature of transformer-based text encoders enabled by padding tokens. The resulting token sequences $\mathbb{V}^N$ are then projected into a continuous vector space through a text embedding function $\varepsilon_t : \mathbb{V}^N \to \mathbb{R}^{N \times d}$, where d denotes the embedding size. For simplicity, we assume that each class label $\mathbf{y}_i$ has already been tokenized, $\mathbf{y}_i \in \mathbb{V}^N$.

A pretrained text encoder $\phi_t : \mathbb{R}^{(N+M)\times d} \to \mathbb{R}^d$ is employed to project a tokenized sequence to a numeric representation of size d. We then define a set P of M learnable soft prompts, $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$.

When extracting the text representation for class k, rather than simply computing $\phi_t(\varepsilon_t(\mathbf{y}_k))$, we prepend the soft prompts to the embedded class label before passing it to the encoder, resulting in:

$$\mathbf{z}_k^t = \phi_t([\mathbf{p}_1, \ldots, \mathbf{p}_M, \varepsilon_t(\mathbf{y}_k)]). \tag{20.1}$$

Since the class labels may vary across different semantic categories (e.g., body parts, medical actions, and pathological conditions), the use of prompt-based conditioning is motivated by the observation that the text encoder might not fully capture the intended semantics of domain-specific terms. By conditioning on learnable prompts, we enable the encoder to adapt its representations in a parameter-efficient manner, allowing it to extract more discriminative features of the target concept.

20.2.3 RDM encoding and classification

Given an input sequence $\mathbf{x}_i$ consisting of T frames of size $H \times W$, a network $\phi_v : \mathbb{R}^{T \times H \times W} \to \mathbb{R}^d$ encodes it to a d-dimensional latent representation as follows:

$$\mathbf{z}_i^v = \phi_v(\mathbf{x}_i). \tag{20.2}$$

This network serves as the visual encoder, replacing the conventional image-based encoder used in CLIP with a 3D convolutional network tailored to process radar-derived temporal sequences.

During training, the visual representation $\mathbf{z}_i^v$ is aligned with its corresponding textual representation $\mathbf{z}_i^t$ through a contrastive loss $\mathcal{L}_{\text{con}}$, which encourages high cosine similarity between matching visual–text pairs and penalizes mismatches within the batch. This objective allows the model to project the RDM sequences into a shared embedding space consistent with their semantic labels.

The overall loss function is defined as:

$$\mathcal{L}(\boldsymbol{\theta}_v, P) = \mathbb{E}_{(\mathbf{x}_i, \mathbf{y}_i) \sim \mathcal{D}} \Big[\mathcal{L}_{\text{con}}(\mathbf{z}_i^v, \mathbf{z}_i^t) \Big] \tag{20.3}$$

where $\mathbf{z}_i^t$ and $\mathbf{z}_i^v$ are obtained using eq. 20.1 and eq. 20.2, respectively.

As shown in Fig. 20.2, classification during inference is performed by computing the cosine similarity between the visual embedding $\mathbf{z}_i^v$ and all textual class embeddings $\mathbf{z}_k^t$, selecting the class with the highest similarity score.

20.3 Experimental evaluation

20.3.1 Dataset

The *MultiMeDaLIS* dataset [241] is a digital collection of LIS with a specific focus on the medical domain. It comprises 126 distinct items, including 100 health-related signs and the 26 letters of the Italian alphabet, designed to support applications in patient-doctor communication contexts. The signs are semantically categorized into body parts, symptoms and conditions, medical procedures, and healthcare-related concepts. They are further organized into semantically marked and unmarked macro-categories, reflecting their relevance and specificity within the medical context.

For this study, we rely exclusively on radar data due to its ability to preserve privacy, capture fine hand motion, and remain unaffected by lighting conditions or the need for physical

contact. The 60 GHz millimeter-wave sensor used (Infineon XENSIV BGT60TR13C) samples at 13 frames per second and captures the three-dimensional movement and positioning of the signer's hands and fingers. It produces time-domain log-scale data from three separate antennas, enabling detailed spatiotemporal analysis of sign articulation.

The dataset includes a total of 25,830 radar sign instances, corresponding to 205 repetitions for each of the 126 signs. Each sign was performed in isolation by a single signer with substantial experience in LIS, starting and ending in a standardized neutral posture.

The resulting dataset offers a vision-free representation of LIS signs in the medical domain, ensuring privacy protection by relying exclusively on radar signals.

20.3.2 Data preprocessing

The raw radar signals acquired from the 60 GHz mmWave sensor undergo a structured preprocessing pipeline to compute the RDM representations. Figure 20.3 presents a simplified overview of this preprocessing pipeline, from raw radar data to final RDM representation.

Each frame consists of time-domain logarithmic-scale data recorded via one transmit antenna and three RX, sampled at 1 MHz with a transmit power level of 31 and an IF gain of 40 dB. Under these parameters, the radar achieves a range resolution of 0.0312 m and a speed resolution of 0.0321 m/s, with a maximum detectable range of approximately 1.60 m and a maximum speed of 4.11 m/s. Each frame is acquired at 0.077 s intervals, centered at 60.50 GHz, allowing for fine-grained capture of hand gestures typical of sign language.

To derive the range-Doppler spectrum, we apply a two-stage FFT process on the time-domain radar data. First, a range FFT is performed over the fast-time (i.e., chirp) dimension after optional mean removal, Blackman-Harris windowing, and zero-padding. The resulting matrix is then transposed, and a Doppler FFT is applied over the slow-time (i.e., chirps) dimension using a Chebyshev window, followed by further zero-padding. The complex 2D output is converted to an absolute magnitude spectrum, and an FFT shift is applied to center the DC component. A configurable spectral thresholding step suppresses low-magnitude values (default of -90 dB, tunable between -75 dB and -120 dB), enhancing signal clarity by reducing background noise.

An optional MTI filter can be applied to further suppress static background signals by zeroing out range-Doppler bins below a predefined velocity threshold. Both the standard and MTI-filtered range-Doppler maps have a resolution of 128×1024. Figure 20.4 illustrates the visual difference between an RDM with and without MTI filtering.

All signal processing steps are designed to preserve the spatiotemporal fidelity of the signing gestures, ensuring that the resulting RDM representations retain the fine-grained

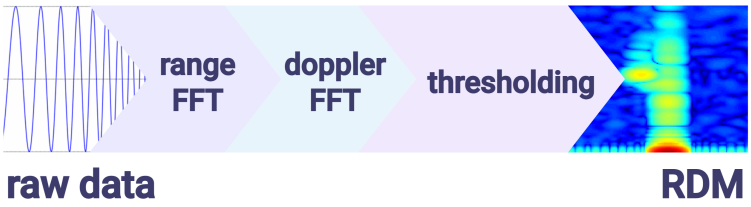


Figure 20.3: Overview of the radar signal preprocessing pipeline leading to the RDM. The diagram shows the main stages of the processing chain, including range FFT, Doppler FFT, and thresholding. While simplified, the figure highlights the key transformation from raw time-domain radar data to a 2D RDM representation. Intermediate operations such as mean removal, windowing, zero-padding, and FFT shifting are omitted for clarity.

motion patterns necessary for accurate recognition.

20.3.3 Metrics

To assess the performance of our sign recognition model, we employ standard classification metrics, including *accuracy*, *sensitivity* (recall), *specificity*, and the F_1 *score*. These metrics are computed from the confusion matrix, which summarizes the counts of true positives, true negatives, false positives, and false negatives. For brevity, the F_1 score is denoted as *F-score* in the tables.

20.3.4 Training procedure

We adopt a 5-fold nested cross-validation, with outer folds for testing and inner folds for validation. The visual encoder is a 3D convolutional backbone, optionally initialized from Kinetics-400. Visual features and text embeddings are projected into a common embedding space, L2-normalized, and compared via scaled cosine similarity, with a learnable logit scale initialized to $\log(1/0.07)$.

We train for 30 epochs with a batch size of 12, minimizing class-balanced cross-entropy with per-class weights inversely proportional to class frequency. Optimization uses AdamW with a learning rate of 0.0001 and weight decay of 0.001. Gradient clipping and automatic mixed precision are enabled, and a plateau scheduler optionally adjusts the learning rate based on validation performance.

20.3.5 Results

To evaluate the effectiveness of our proposed method, we compare it against a range of state-of-the-art baselines, including RGB, RGB-D, and radar-based sign language recognition

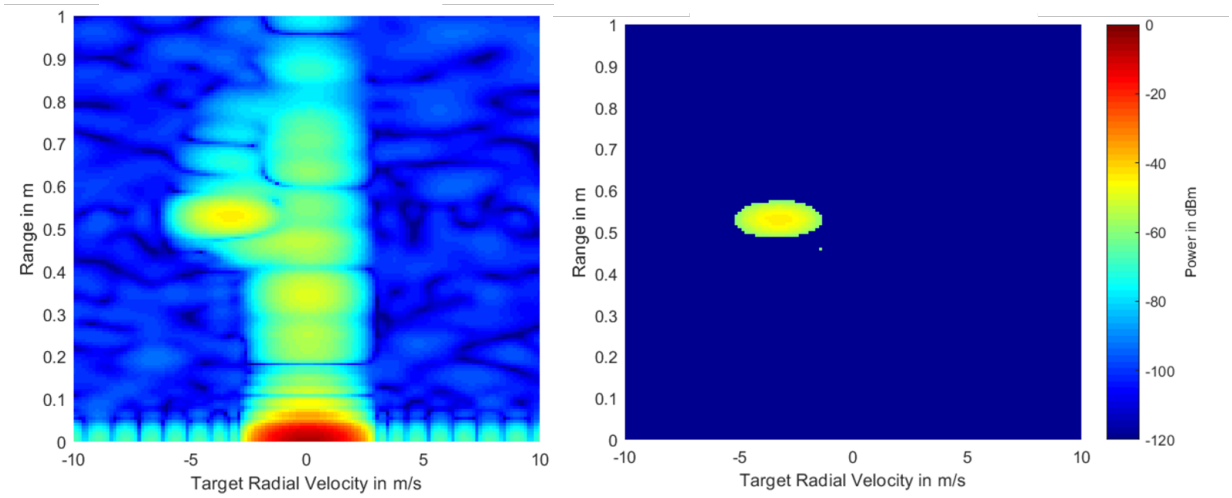


Figure 20.4: RDM with and without MTI filtering. Comparison between the same RDM without MTI filtering (left) and with MTI filtering applied (right). The MTI filter suppresses static background signals by zeroing out bins below a predefined velocity threshold, resulting in a cleaner representation of dynamic hand motion.

Method	**Data type**	**Acc.**	**Sens.**	**Spec.**	**F-score**
De Coster et al. [76]	RGB	88.4 ± 3.7	78.6 ± 3.4	98.2 ± 3.5	79.7 ± 3.9
Vahdani et al. [368]	RGB-D	84.1 ± 2.9	83.1 ± 4.1	85.1 ± 2.1	84.2 ± 2.2
Caligiore et al. [41]	RGB-D	74.6 ± 3.9	66.4 ± 3.1	82.8 ± 3.0	59.9 ± 3.3
Jhaung et al. [161]	Radar	71.9 ± 3.3	61.9 ± 2.3	81.9 ± 3.4	69.2 ± 2.9
Debnath et al. [81]	Radar	79.3 ± 3.5	83.3 ± 3.3	75.3 ± 3.1	79.3 ± 3.9
Arab et al. [12]	Radar	81.0 ± 3.4	65.9 ± 3.1	96.1 ± 2.8	29.6 ± 2.9
Mineo et al. [240]	3×RDM + 3×MTI	93.6 ± 0.5	87.9 ± 0.5	99.3 ± 0.6	87.7 ± 0.5
TRACE (ours)	**3×RDM + 3×MTI**	**99.3** ± 0.1	**98.8** ± 0.2	**99.9** ± 0.1	**98.8** ± 0.2

Table 20.1: Comparison with state-of-the-art methods in sign language recognition. The "Data type" column indicates the primary sensor modality employed by the method. The best performance is achieved by TRACE, which combines 3×RDM and 3×MTI inputs, highlighting the importance of multimodal radar encoding and multi-perspective information. TRACE leverages the MixConv architecture pretrained on radar data and employs 64 soft prompts, yielding the best overall performance among all methods.

approaches. For methods without publicly available code, we implemented the models ourselves based on the original descriptions, reproducing them to the best of our ability. All models were evaluated using their default hyperparameters, with model selection guided by validation performance. The results are reported in Table 20.1, using the evaluation metrics described.

Among existing approaches, RGB-based and RGB-D-based models often achieve strong performance in isolated sign recognition tasks. However, these methods rely on visual input, which may raise privacy concerns in sensitive environments such as hospitals or rehabilitation centers. Moreover, their performance can degrade under challenging conditions such as variable lighting or cluttered backgrounds, limiting their applicability in real-world scenarios.

Despite being privacy-preserving, prior radar-based methods [161, 81, 12] have generally underperformed compared to their vision-based counterparts. This performance gap can be attributed to the use of handcrafted features or shallow architectures, which fail to capture the rich spatiotemporal structure of sign articulation. Jhaung et al. [161] and Arab et al. [12] report F_1 scores below 70%, while Debnath et al. [81] achieves 79.3% by incorporating additional radar features.

Our method, which combines three antenna inputs with both RDM and MTI representations, overcomes these limitations and achieves superior performance across all metrics. The integration of RDM and MTI provides richer motion cues that, together with spatiotemporal modeling and prompt-conditioned contrastive learning, enhance the model's ability to capture sign dynamics. This is further supported by a comparison with Mineo et al. [240], who use the same radar dataset. While their method achieves strong overall results, TRACE achieves higher performance, demonstrating the benefits of multimodal radar encoding and language-aligned supervision.

20.3.6 Ablation study

Data type	**Acc.**	**Sens.**	**Spec.**	**F-score**
RDM	95.2 ± 0.4	94.6 ± 0.2	95.9 ± 0.7	93.8 ± 0.2
3×RDM	97.9 ± 0.3	97.5 ± 0.4	98.3 ± 0.2	97.9 ± 0.1
MTI	92.0 ± 0.3	90.4 ± 0.8	93.6 ± 0.3	90.4 ± 0.5
3×MTI	92.4 ± 0.2	90.6 ± 0.4	94.3 ± 0.1	91.3 ± 0.8
RDM + MTI	97.8 ± 0.4	97.4 ± 0.6	98.2 ± 0.2	97.3 ± 0.3
3×RDM + 3×MTI	**99.3 ± 0.1**	**98.8 ± 0.2**	**99.9 ± 0.1**	**98.8 ± 0.2**

Table 20.2: Impact of radar input types and number of antennas on model performance. We evaluate the effect of using RDM, MTI, and their combination, in both single-antenna and multi-antenna configurations. The best performance is achieved by combining 3×RDM and 3×MTI inputs, highlighting the importance of multimodal radar encoding and multi-perspective information.

To assess the impact of each component in our model, we perform a series of experiments focusing on three key aspects: radar input configuration, backbone architecture, and the number of soft prompts used in textual supervision.

Radar input types and antenna configuration.

Table 20.2 evaluates the effect of radar modality and antenna configuration, comparing RDM, MTI, and their combination under both single-antenna and multi-antenna setups. Results show that combining RDM and MTI outperforms single-modality inputs, as the two provide complementary information: RDM encodes spatial and velocity cues, while MTI focuses on motion saliency. In addition, multi-antenna input offers diverse perspectives that enhance robustness. The best results are obtained using 3×RDM and 3×MTI, confirming the effectiveness of both modality fusion and multi-view radar encoding.

Architecture	**Acc.**	**Sens.**	**Spec.**	**F-score**
ResNet 3D	93.5 ± 0.7	87.9 ± 0.4	99.0 ± 0.9	90.3 ± 0.4
↪ Pre-tr.	98.0 ± 0.1	96.1 ± 0.2	99.9 ± 0.1	96.8 ± 0.6
R(2+1)D	88.1 ± 2.8	85.3 ± 3.4	90.9 ± 2.2	87.4 ± 2.6
↪ Pre-tr.	97.5 ± 0.4	95.0 ± 0.8	99.9 ± 0.1	95.9 ± 0.3
MixConv	93.2 ± 0.6	86.8 ± 0.9	99.6 ± 0.3	85.7 ± 2.0
↪ **Pre-tr.**	**99.3 ± 0.1**	**98.8 ± 0.2**	**99.9 ± 0.1**	**98.8 ± 0.2**

Table 20.3: Effect of backbone architecture and pretraining on performance. We compare ResNet 3D [134], R(2+1)D [364], and MixConv [354] with and without pretraining on radar data. Results show that both architectural choice and pretraining substantially impact performance, with the pretrained MixConv achieving the highest scores across all metrics.

Backbone architecture and pretraining.

In Tab. 20.3, we compare three backbone architectures: ResNet 3D [134], R(2+1)D [364], and MixConv [354], each evaluated with and without pretraining on radar data. The results show that pretraining consistently improves performance, highlighting the value of prior exposure to radar-specific spatiotemporal patterns. Among the tested models, pretrained MixConv achieves the best overall results, suggesting that its mixed convolutional structure is particularly effective in capturing the heterogeneous motion dynamics of sign language.

Prompt-based textual supervision.

Table 20.4 reports the effect of varying the number of soft prompts used in the text encoder. Performance consistently improves as the number of prompts increases, with only marginal gains

# prompts	Acc.	Sens.	Spec.	F-score
0	90.3 ± 0.2	86.3 ± 0.7	94.4 ± 0.3	89.0 ± 0.3
4	92.8 ± 1.1	89.1 ± 0.7	96.4 ± 2.8	92.9 ± 0.3
8	93.6 ± 1.4	93.8 ± 1.2	93.3 ± 3.9	93.0 ± 0.8
16	96.6 ± 0.1	94.3 ± 1.2	98.8 ± 1.1	95.2 ± 0.3
32	97.8 ± 0.6	97.0 ± 0.2	98.6 ± 1.0	98.5 ± 0.3
48	98.5 ± 0.5	98.2 ± 0.2	98.7 ± 0.7	98.3 ± 0.4
64	**99.3 ± 0.1**	**98.8 ± 0.2**	**99.9 ± 0.1**	**98.8 ± 0.2**

Table 20.4: Effect of the number of soft prompts on classification performance. We evaluate the effect of varying the number of learnable prompts prepended to each class label. Results show that using 64 prompts leads to the best performance across all metrics, highlighting the benefit of richer prompt-based conditioning for textual supervision.

at 4 and 8 but substantial improvements from 16 onwards. The best performance across all metrics is achieved with 64 prompts. These results support our hypothesis that, since class labels span heterogeneous semantic categories, the frozen text encoder alone may not fully capture their intended meaning. Prompt-based conditioning introduces additional learnable context that helps disambiguate and align domain-specific terms with radar features, thereby enhancing classification performance.

20.4 Discussion

TRACE frames radar-based isolated SLR as a language-grounded representation learning problem, where spatiotemporal motion cues extracted from Range-Doppler sequences are aligned with textual class semantics via prompt-conditioned contrastive learning. This design is motivated by two practical constraints that frequently limit camera-based deployments in clinical and assistive settings: (i) privacy constraints that restrict the capture or storage of identifiable visual data, and (ii) acquisition variability (lighting, background clutter, partial occlusions) that can degrade the stability of vision-first pipelines. By operating on radar-derived motion representations, TRACE preserves privacy by construction while retaining robustness to environmental factors that are not directly relevant to signing.

Empirically, the evaluation on *MultiMeDaLIS* indicates that the proposed alignment strategy yields a consistent advantage over prior baselines, including radar-centric models and an RGB-based reference. The results suggest that the performance gains are not attributable to a single component, but rather to the interaction of three ingredients: multimodal radar representations that preserve complementary motion information, pretrained spatiotemporal encoders that provide a strong inductive bias for sequence modeling, and prompt-conditioned textual supervision that shapes the

embedding space toward semantically meaningful structure. In particular, contrastive alignment acts as a regularizer that discourages purely dataset-specific shortcuts, while learnable soft prompts support label representations that better match the target domain than fixed templates, improving discriminability among visually and kinematically similar signs.

From a deployment perspective, these findings strengthen the case for radar as a viable sensing modality for sign language understanding in sensitive environments. Radar-based acquisition reduces the exposure of identity-revealing details and can support more acceptable clinical workflows, especially when combined with on-edge inference and minimal retention policies. At the same time, the approach retains a practical inference rule, nearest text embedding, that naturally supports interpretability at the label level and is compatible with incremental extension of the vocabulary through text-conditioned representations, subject to validation.

A limitation of the current formulation is that TRACE addresses isolated sign recognition and does not explicitly model temporal segmentation boundaries. This restricts direct deployability for full-sentence clinical translation, where the system must detect when signs start and end, handle variable signing rates, and interpret transitions between consecutive signs. In continuous signing, boundaries are often ambiguous because of coarticulation, and the meaning of a sign may depend on surrounding context and non-manual components. Therefore, applying the current isolated-sign classifier frame by frame would not be sufficient for realistic patient-doctor dialogue. This limitation is particularly relevant in medical communication, where negation, symptom duration, pain intensity, and temporal references may depend on sentence-level context rather than on individual lexical units alone.

This design choice was intentional at the present stage: the objective of this chapter is to evaluate whether radar-derived representations can support privacy-preserving recognition of medical signs under controlled conditions. Isolated recognition provides a necessary first step because it allows the sensing modality, radar preprocessing, and text-aligned representation to be assessed without the additional confounder of temporal segmentation. The next step is the construction of a continuous sign language recognition dataset for the medical domain, including sentence-level recordings, explicit or weak temporal boundary annotations, and evaluation protocols for segmentation and sequence prediction. Such a dataset would enable models that jointly learn boundary detection, sign recognition, and semantic grounding, moving the approach toward full-sentence clinical translation in realistic interaction scenarios.

Several directions emerge as priorities for future work. First, extending TRACE from isolated to **continuous** sign recognition requires addressing segmentation, coarticulation, and streaming inference with bounded latency, potentially through causal temporal modules and decoding strategies that remain stable under variable signing pace. Second, **cross-lingual generalization** across different sign languages is an appealing direction, but it introduces challenges related to lexical mismatch, culturally grounded semantics, and differing articulation patterns; here, prompt design and label semantics will likely require careful adaptation and evaluation protocols. A third priority

is broader robustness analysis, including cross-signer and cross-setup testing, to quantify how radar placement, environmental reflections, and motion regimes affect the learned embedding space and to identify failure modes that are not visible in controlled acquisition.

Overall, TRACE supports a broader conclusion: language supervision can be used not only as an output layer for vision-language models, but as a training signal that meaningfully organizes radar embeddings for privacy-aware sign recognition. This makes radar-based SLR a realistic candidate for clinical and assistive communication scenarios where privacy and deployment constraints are binding, while also opening a path toward more scalable vocabularies and more natural interaction settings through future extensions.

20.5 CO_2 Emissions Related to Experiments

Experiments were conducted using a private infrastructure, with an estimated carbon efficiency of 0.432 kg CO_2eq/kWh. A cumulative of 290 hours of computation was performed on four NVIDIA A100 GPUs (thermal design power of 300 W), with total emissions estimated[1] to be 31.32 kg CO_2eq.

[1]Estimations were performed using the Machine Learning Impact calculator [189].

Chapter 21

Impact and Community Contribution: The SignEval 2025 Challenge at the ICCV Multimodal SLR Workshop

Source note. The following chapter is based on and adapted from:

- Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, Jihwan Moon, Václav Javorek, et al. The signeval 2025 challenge at the iccv multimodal sign language recognition workshop: Results and discussion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2025

21.1 Overview

SLR aims to reduce communication barriers between Deaf or hard-of-hearing individuals and non-signers by automatically interpreting signed content. Unlike spoken languages, sign languages are expressed through tightly coupled manual cues (hand shape, orientation, movement) and non-manual cues (facial expressions and body posture), making recognition inherently multimodal and temporally complex [91, 337]. Within SLR, two problem settings are commonly distinguished: ISLR, which classifies short clips corresponding to single signs [196], and CSLR, which transcribes full sentences into gloss sequences. CSLR is particularly challenging because it must handle long-range temporal dependencies, coarticulation, and compositional linguistic structure [7]. Despite recent progress, key obstacles remain, including signer variability, modality constraints, and the need for scalable systems that generalize outside training conditions [305].

Most SLR pipelines still rely on RGB video inputs, but alternative representations are gaining

attention. In particular, pose representations, encoding body keypoints, provide a compact abstraction of signing behavior that can help mitigate signer overfitting, support signer privacy preservation, improve cross-domain robustness, and enable lightweight computation [150, 5, 149, 167]. However, pose-based SLR remains comparatively underexplored relative to other modalities, motivating standardized benchmarks that isolate generalization challenges and make progress comparable across methods.

To advance the field with reproducible evaluation and to encourage research beyond vision-only settings, the *SignEval 2025* challenge (organized at ICCV 2025 within the Multimodal Sign Language Recognition workshop) introduced two complementary tracks spanning CSLR and ISLR. The first track targets CSLR using the newly curated *Isharah* dataset [6], a Saudi Sign Language resource comprising over 14,000 RGB videos corresponding to approximately 1,000 unique sentences signed by 18 individuals. Participants were provided with 2D pose modality data extracted from RGB videos via MediaPipe, prioritizing structure, scalability, and signer invariance. Two tasks were defined: *Signer-Independent*, which evaluates cross-signer generalization (a core requirement for deployable CSLR), and *Unseen-Sentences*, which evaluates generalization to novel sentence compositions by leveraging learned grammar and semantics.

The second track targets ISLR using *MultiMeDaLIS* [41], a multimodal dataset designed for doctor-patient communication in Italian Sign Language (LIS, Lingua Italiana dei Segni). *MultiMeDaLIS* provides 25,830 isolated sign instances across 126 classes, captured synchronously with three complementary privacy-preserving modalities: 60 GHz radar, RGB video, and skeletal data. This setting enables systematic study of ISLR beyond vision-only pipelines, fostering research into modality robustness, sensor fusion, and clinically motivated vocabularies. In the challenge, participants were asked to classify isolated signs using only radar and RGB modalities, and results were reported through two public leaderboards that established new benchmarks across both tracks.

This paper summarizes the outcomes of *SignEval 2025*, detailing the datasets and tasks, the evaluation setup, participant submissions, and the best-performing approaches for both CSLR and ISLR. By jointly benchmarking signer generalization, linguistic generalization to unseen sentences, and multimodal recognition under privacy-aware sensing, the challenge provides a first consolidated reference point for multimodal SLR and highlights open problems that motivate future research. Additional resources (tasks, leaderboards, baselines, and development kits) are available at https://multimodal-sign-language-recognition.github.io/ICCV-2025/.

21.2 Track 1: Continuous Sign Language Recognition Track

The first track of the SignEval 2025 challenge is a shared task on CSLR, focusing on recognizing gloss sequences (i.e., sign language sentences) from pose-based input streams. Participants are provided

with pose modality data derived from the Isharah *dataset* [6] and are challenged to develop models that can effectively generalize to unseen signers and novel sentence structures. For this task, a subset of the Isharah dataset was used, consisting of over 14,000 RGB videos covering approximately 1,000 unique sentences in Saudi Sign Language (SSL), performed by 18 different signers.

Two shared tasks have been organized within this track: signer-independent and unseen sentences recognition tasks.

Task 01, Signer-Independent. In this task, models are trained on a subset of signers and evaluated on entirely unseen signers. This tests the model's ability to generalize across individuals, a critical property for scalable real-world CSLR systems. The training set for this task consists of 10,000 samples performed by 12 signers. The development set includes 950 samples performed by one signer who does not appear in the training data. The test set contains 3,800 samples performed by five signers who are unseen in both the training and development sets.

Task 02, Unseen-Sentences. This task evaluates the model's ability to recognize sign sequences that were not seen during training. Although the model has been exposed to the individual glosses in various contexts during the training, it has never encountered them in the specific sentence structures used for testing. The dataset used for this task consists of 9,900, 550, and 550 samples for the training, development, and test sets, respectively.

21.2.1 Dataset

The CSLR tasks leverage the Isharah dataset [6], which is a large-scale, multi-scene resource developed to advance research in CSLR and sign language translation. It contains 30,000 video clips of SSL sentences recorded by 18 proficient SSL users, including deaf individuals, hard-of-hearing participants, and certified interpreters. The videos capture 2,000 unique sentences across domains such as healthcare, transportation, education, legal services, and emergency scenarios. These recordings were made by participants using smartphones in natural environments, introducing a wide range of visual variability that mirrors real-world conditions.

For the CSLR shared tasks, we use the Isharah-1000 subset, which consists of 14,000 videos and 1,000 unique sentence types. To extract the pose data from the dataset, the MediaPipe Holistic model [411] is utilized, which estimates full-body, hand, and facial landmarks from each video frame. We select 86 pose keypoints consisting of 21 landmarks from each hand, 24 from the upper body, and 20 keypoints representing the lip contour. The dataset used for the *Signer-Independent* task was split into 10,000, 950, and 3,800 samples for training, development, and testing sets, respectively. Each set contains signers not present in the others. The dataset used for the *Unseen-Sentences* task consists of 9,900, 550, and 550 samples for the training, development, and test sets, respectively.

21.2.2 Evaluation Metrics

For the CSLR challenge tasks, we employ the *Word Error Rate (WER)* as the primary metric to evaluate system performance. WER measures the proportion of errors in the recognized (hypothesized) gloss sequence relative to the reference (ground-truth) gloss sequence, accounting for substitutions (Sub.), deletions (Del.), and insertions (Ins.) It is defined as:

$$\text{WER} = \frac{Sub. + Del. + Ins.}{ReferenceLength} \times 100\% \tag{21.1}$$

21.2.3 Results and Participating Teams

Baselines

We proposed a baseline for both tasks of the CSLR track using 2D keypoints extracted from hand landmarks. Each frame is represented by 42 keypoints from both hands, resulting in a sequence tensor $P = \{p_1, p_2, \ldots, p_T\}$ of shape $T \times 42 \times 2$, which is flattened to $T \times 84$ and processed through a linear projection layer into a d-dimensional hidden space. The proposed baseline adopts an architecture built upon the standard Transformer encoder [373], enhancing temporal modeling through stacked attention layers and temporal pooling. It consists of four Transformer encoder blocks with sinusoidal positional encodings to preserve temporal structure. No intermediate supervision is applied between these blocks, allowing residual accumulation of temporal context. After the Transformer layers, the model performs progressive temporal pooling using two 1D average pooling layers interleaved with temporal convolutional layers to reduce sequence length and capture broader contextual patterns. A final multi-layer perceptron outputs per-frame gloss predictions.

Results

The results of the two tasks in this track are presented in Table 21.1 (Signer-Independent task) and Table 21.2 (Unseen-Sentence task). Across both tasks, the top three teams converged on two primary paradigms: Conformer-based encoders and Graph Convolutional Networks (GCNs). Conformer-based encoders interleave convolutional modules for fine-grained and local feature extraction with self-attention layers for capturing long-range dependencies. The GCNs jointly process semantically grouped 3D keypoints (e.g., face, hands, body) in parallel streams before fusing their representations.

Each team conducted a detailed analysis of the Isharah dataset to tailor their architectures and auxiliary modules to the distinct challenges posed by signer variability (Task 1) versus the greater demands of unseen-sentence generalization (Task 2). Almost all solutions adopted a *hybrid attention* strategy using global self-attention to model long-range temporal context and local convolutions or learnable pooling for adaptive downsampling and refinement of short-range temporal patterns.
Notably, Task 2 proved substantially more challenging, where WERs roughly tripled compared to Task 1 for most teams, underscoring the difficulty of generalizing to novel sentence structures

without any prior exposure during training.

As shown in Tables 21.1 and 21.2, a clear pattern of overfitting emerges when comparing the WERs of the development and test sets. Some teams that achieve very low WER on the development set often incur substantially higher Test WER, indicating models have memorized idiosyncrasies of the held-out development data rather than learning robust generalizations. For example, in Task 1, the VIPL_SLP team achieved a development WER of 25.16% and a test WER of 7.45%, which indicates a surprisingly better performance on unseen test data. This can be attributed to the use of careful regularization and data augmentation strategies that effectively mitigated overfitting. By contrast, the CV Group (UWB) exhibits a development WER of 48.42% whereas their test WER is 33.82%. This inverted gap points to potential data leakage or overly aggressive model pruning tuned on the development set.

In Task 2, the overfitting gap widens considerably, where the CV Group (UWB) reported a development WER of just 2.77%, yet suffers an 80.27% WER on Test, a more than 28× increase that typifies extreme over-adaptation to development sentences. Similarly, the development WER of the Decoder team is 9.44%, whereas their test WER jumpted significantly to 107.27%. The increase confirms that heavily parameterized models can *memorize* development transcripts but fail to generalize to novel sentence structures without stronger regularization or cross-validation rigor.

Overall, the discrepancies between development and test WERs underscore the need for robust regularization and more diverse development folds. Moreover, CSLR models should focus on learning the sign gestures rather than memorizing signer-specific patterns, in order to enhance robustness to signer variation.

#	**Team Name**	**Dev WER (%)**	**Test WER (%)**
1	VIPL_SLP [400]	25.16	07.45
2	Kronus [366]	09.90	09.65
3	Decoder [92]	05.62	12.01
4	CPAMI [133]	07.31	13.07
5	Kumass2020	06.94	16.67
6	RW [169]	15.37	20.45
7	StarAtNyte*	–	31.11
8	CV Group (UWB)	48.42	33.82
9	Baseline	25.10	38.50

Table 21.1: Results for Task 1 (Signer-Independent) in SignEval 2025 Challenge Track 1, showing both development-set and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams.

#	Team Name	Dev WER (%)	Test WER (%)
1	VIPL_SLP [400]	36.90	28.20
2	CPAMI [133]	55.08	47.78
3	Kronus [365]	–	55.28
4	Kumass2020	62.87	57.68
5	CV Group (UWB)	2.77	80.27
6	Baseline	87.60	81.20
7	Decoder [92]	9.44	107.27
8	Wonjune.kim*	108.63	146.00
9	RW [169]	140.95	283.56

Table 21.2: Results for Task 2 (Unseen Sentence) in MSLR Challenge Track 1, showing both development-set and test-set WERs. * indicates the team did not submit a report and is not listed among participating teams.

Participating Teams

VIPL_SLP [400] proposed a skeleton-based CSLR framework using a unified CNN/GCN-Conv1D-BiLSTM pipeline. Based on CoSign [167], keypoints are grouped (e.g., hands, face, body) and processed by group-specific GCNs. Alternatively, they encoded keypoints as Gaussian heatmaps, processed via CNNs. Frame-level features are aggregated with 1D convolutions, followed by a two-layer BiLSTM for temporal modeling. Predictions from both local (frontend) and global (backend) features are supervised with CTC loss. Motion cues from joint-wise offsets and regularization enhance the fusion of multi-stream signals. Models were trained on 86 keypoints. Their single-stream model consisted of three-layer group-wise GCNs as feature extractors, followed by 1D convolutions and a BiLSTM with 1024 hidden units for local and global temporal modeling, respectively.

Kronus presented two complementary pose-based architectures, both leveraging 2D pose keypoints and anatomically dividing the input into four regions: right hand, left hand, face/lips, and upper body/torso. Each region is processed independently using a GCN-based stream, followed by either temporal compression or an MLP, and their features are fused for downstream decoding. The first model [366], targeting Signer-Independent CSLR, uses a Transformer encoder and a permutation-trained autoregressive decoder, with auxiliary gloss decoders supervising each region. The second model [365], designed for generalization to unseen sentence compositions, employs a two-stage temporal convolutional backbone, with a shared classifier producing gloss predictions at multiple temporal resolutions, all trained using CTC losses. Both models incorporate auxiliary supervision on region streams to ensure meaningful contributions from all parts, enhancing robustness to signer variability and sentence complexity.

Decoder [92] proposes a framework designed for signer-independent recognition, leveraging anatomically partitioned 2D pose sequences. The input skeleton is divided into four regions: right hand, left hand, lips, and torso, each encoded independently using region-specific GCNs followed

by temporal compression via 1D convolutions. These part-wise features are concatenated and processed through a Transformer encoder. Gloss sequences are generated using a permutation-trained autoregressive Transformer decoder that models flexible gloss dependencies beyond monotonic left-to-right order. Auxiliary gloss decoders supervise each region stream to enforce semantic alignment and improve robustness, particularly in underrepresented regions. This architecture balances detailed local motion modeling with global temporal context to handle the complexities of continuous sign language.

CPAMI [133] addressed the challenges of signer variability and syntactic generalization by developing two specialized pose-based architectures, each tailored to the task setting. For the Signer-Independent task, they employed a Conformer-based model that combines convolutional and self-attention mechanisms to capture both local and global temporal patterns. The architecture begins with a shallow temporal encoder using 1D convolutions, followed by stacked Conformer blocks that integrate multi-head self-attention, depthwise convolutions, and feed-forward layers. This hybrid modeling approach yields signer-agnostic representations suitable for CTC-based gloss prediction. For the US track, they proposed a Multi-Scale Fusion Transformer architecture featuring a joint attention mechanism and a novel dual-path temporal encoder. This encoder processes input sequences through both fine-grained and downsampled branches, fusing them into a rich multi-scale representation that is then passed through a Transformer encoder and an MLP-based classifier head. Both systems use CTC loss for end-to-end training without frame-level alignment.

Kumass2020 proposed a Temporal Pose-aware CSLR (TemPo) model to learn structured spatiotemporal representations from 3D pose keypoints. The input to TemPo is a tensor $X \in \mathbb{R}^{T \times D}$, where T is the number of frames and D is the pose vector dimensionality. The architecture includes a group-wise embedding that processes semantic groups of the human pose (e.g., face, hands, body) via dedicated MLPs, producing disentangled feature representations. These embeddings are concatenated and passed through a hybrid backbone combining Transformer encoder layers for capturing global context and TCNs with a channel-wise learnable pooling layer for local feature refinement and adaptive temporal downsampling. To improve robustness and reduce overfitting, the team employed a Semantic Consistency Regularization (SCR) strategy involving structured masking of semantic keypoint groups during training, encouraging consistent predictions through a symmetric KL divergence loss alongside standard CTC losses. During inference, the full pose sequence is used without masking, and CTC decoding produces the final gloss sequence.

RW team [169] proposed a decoder-only autoregressive architecture for continuous sign language recognition, treating the task as a conditional gloss generation problem. Their model, named AutoSign, adapts the DTrOCR framework [109] to generate Arabic gloss sequences from 2D skeleton data extracted using pose estimation methods. The input pose sequence is grouped into four semantic parts (left/right hand, face, body), then processed with part-aware augmentations before being compressed by a two-layer 1D-CNN. This compressor reduces the temporal resolution and increases the feature dimensionality from 134 to 512. Temporal order is preserved via learnable

positional embeddings. The processed pose features are projected and concatenated with gloss token embeddings, and passed through a decoder-only transformer based on AraGPT2 [11], which uses causal and cross-attention to autoregressively predict gloss tokens. The system is trained using cross-entropy loss over non-padded gloss tokens.

CV Group (UWB) focused primarily on preprocessing efforts. Their approach involved normalising signer-specific keypoints, pruning inconsistent annotations, and generating gloss-level labels suitable for training. Two primary approaches were evaluated: a customised T5-based model adapted from prior research [296], and Uni-Sign [204], a recently proposed pre-trained architecture designed for Sign Language Recognition. Both methods incorporated Sign Space normalization and augmentation strategies described in [296].

21.3 Track 2: Isolated Sign Language Recognition

21.3.1 Task Description

The ISLR Challenge seeks to advance multimodal sign-language and gesture recognition by providing a comprehensive dataset that pairs RGB video with 60 GHz mm-wave RDM measurements. Its structured yet flexible competition format invites teams to devise and evaluate novel methods that harness the complementary strengths of these modalities and push the boundaries of multimodal fusion. By leveraging the newly released *MultiMeDaLIS* dataset [241], the challenge encourages systems able to capture subtle hand shapes, intricate motion dynamics, and fine-grained temporal cues, setting a new benchmark for multimodal understanding.

21.3.2 Dataset

MultiMeDaLIS [241] is an LIS dataset tailored to patient–doctor communication. It comprises 126 lexical items (100 healthcare-specific signs and the 26 letters of the Italian one-hand manual alphabet) each performed 205 times by a proficient signer, yielding 25,830 annotated instances. Data were recorded synchronously with four privacy-preserving, non-invasive sensors: a three-antenna 60 GHz radar array (13 fps), 720p 30 fps RGB-D video with facial-landmark tracking (Intel RealSense D455), dedicated facial-expression streams from a Kinect v1, and 1080p 25 fps depth point clouds from a Zed 2. This multimodal capture spans both manual articulators and subtle non-manual cues (e.g., facial movements, torso shifts, mouthing), yielding a dataset larger and more suitable for supervised learning than previously available radar-based or RGB-D-based sign-language corpora. The dataset is partitioned into 12,600 labelled training samples across 126 gesture-specific directories, and 4,410 validation samples plus 7,560 test samples stored in class-agnostic directories with concealed labels. Each sample directory contains one frontal RGB video and three temporally aligned range-Doppler clips, preserving multimodal synchrony while masking class information in the file paths.

21.3.3 Evaluation Metrics

For the ISLR track, we evaluate the performance of the recognition models using the standard classification metrics: *accuracy*, *sensitivity* (recall), *specificity*, and the F_1*-score*. All these metrics are derived from the confusion matrix, which comprises true positives, true negatives, false positives, and false negatives.

21.3.4 Results and Participating Teams

Baselines

To contextualize the results achieved by participants in the challenge, we define a set of reference baselines that include both standard deep learning architectures and recent domain-specific approaches relevant to LIS recognition in medical contexts.

As a core benchmark, we adopt the ResNet(2+1)d architecture as in Caligiore et al. [41], which extends standard 2D convolutional networks by factorizing 3D spatiotemporal filters into separate spatial and temporal components. This architecture is well-suited to capture motion dynamics in video-based sign language recognition while maintaining a balanced computational cost. We evaluate ResNet(2+1)d under three input modalities: RGB only, RGB combined with depth (RGB-D), and RGB combined with Range-Doppler Maps (RGB+RDM), the latter incorporating radar-derived motion information into the visual stream.

Moreover, Mineo et al. have recently introduced two radar-only architectures for clinical LIS: one leveraging a residual autoencoder on Range-Doppler Maps with a subsequent Transformer for classification [240], and another (TRACE) combining a ResNetMC backbone with CLIP-based text alignment to directly associate radar features with natural-language labels, thereby improving semantic coverage of medical sign vocabularies [249].

Results

Table 21.3 reports performance on the ISLR challenge track. The R(2+1)d model from Caligiore et al. [41] serves as a visual baseline, achieving 71.8% accuracy on RGB alone and 74.6% when incorporating depth. The two radar-only approaches by Mineo et al. [240, 249] demonstrate the power of RDM+MTI inputs, both reaching 93.6% accuracy, which far exceeds the RGB-only baseline and confirms the efficacy of radar for privacy-preserving ISLR.

Among the challenge participants, fusion of RGB+RDM modalities yields near-perfect scores: *ROLL TIDE RADAR* and *Former Human* both achieve 99.8% accuracy, while *IMPRESS NCSU* and *CPAMI (UW)* closely follow at 99.7% and 99.4%, respectively. These results underline the advantage of multimodal integration, particularly late fusion strategies, in leveraging complementary visual and motion cues. Notably, leaderboard entries such as *ML* and *SabyasachiBiswas147* maintain similarly high accuracy (99.3% and 97.1%), whereas *Matteo Cacioppo* submission reaches 78.1%,

	Data type	Acc.	Sens.	Spec.	F1-Sc.
R(2+1)d [41]	RGB	71.8	64.5	79.1	58.2
R(2+1)d [41]	RGB-D	74.6	66.4	82.8	59.9
Mineo et al. [240]	RDM+MTI	93.6	87.9	99.3	87.7
TRACE [249]	RDM+MTI	99.3	98.8	99.9	98.8
R(2+1)d [240]	RGB+RDM	79.4	70.2	88.6	75.5
ROLL TIDE RADAR [323]	RGB+RDM	99.8	99.9	100.0	99.9
Former Human [172]	RGB+RDM	99.8	99.8	100.0	99.8
IMPRESS NCSU [226]	RGB+RDM	99.7	99.7	100.0	99.7
CPAMI (UW) [155]	RGB+RDM	99.4	99.4	100.0	99.4
ML	RGB+RDM	99.3	99.3	100.0	99.4
SabyasachiBiswas147	RGB+RDM	97.1	97.1	100.0	97.1
Matteo Cacioppo	RGB+RDM	78.1	78.1	99.8	78.9

Table 21.3: Baseline and Leaderboard for the SignEval Challenge Track 2.

suggesting that model architecture and fusion design critically impact final performance. Overall, the top-ranked teams substantially outperform both the vision-only and radar-only baselines, confirming that joint RGB–radar systems offer the most robust solution for isolated sign recognition in clinical contexts.

Participating Teams

ROLL TIDE RADAR [323] combines a state-of-the-art video–based ISLR model with a novel radar envelope pipeline. For the RGB branch, they adopt Uni-Sign, a dual-encoder architecture that fuses a vision encoder and a pose encoder via a prior-guided fusion module before querying mT5-Base for classification, fine-tuned on *MultiMeDaLIS* with an Italian-language head. For the radar branch, they extract kinematic envelopes (upper and lower) from single-antenna RDMs to condense the 3D data cube into paired 2D heatmaps, which are then resized and fed into pretrained image classifiers (ConvNeXtV2, Swin Transformer V2, DeiT, and VGG16) in order to form RDM-Env. The Uni-Sign branch uses the publicly released WLASL100 weights and is fine-tuned for 20 epochs with AdamW (learning rate 3×10^{-4}), a batch size of 8, and a cross-entropy loss on pose-processed clips from RTMPose. The radar envelope models are trained by first computing energy-threshold envelopes (95% upper, 5% lower) per frame, stacking them into $2 \times T \times R$ matrices, and generating RGB-style heatmaps. Each candidate classifier is then fine-tuned on these 2D inputs for 20 epochs using AdamW at 3×10^{-4}, batch size 8, and standard ImageNet normalization. Performance is monitored via validation accuracy, with the best model selected for late fusion.

Former Human [172] implements a two-branch architecture. The RGB branch is a 3D-ResNet–18 model pretrained on Kinetics-400 and fine-tuned to the 126-class vocabulary by replacing its final fully-connected layer with a 126-unit linear head. The Radar branch employs a diffusion-based U-Net trained to reverse Gaussian noise over 400 timesteps on RDMs. Latent features are extracted from the fourth down block and fed to one of three classifier heads (linear, GRU, or partially

fine-tuned U-Net + spatial-temporal head) to produce 126 logits. The RGB branch was trained for 15 epochs on a single NVIDIA P100 GPU using AdamW with a learning rate of 1×10^{-4} and default weight decay. They processed 16-frame RGB clips, resized to 224×224 and normalized with ImageNet statistics, in batches of eight. Cross-entropy loss was optimized while logging metrics to TensorBoard and a CSV file at each iteration; they saved both the final and best-performing model weights. For the radar branch, three strategies were explored: first, a linear head trained with AdamW (1×10^{-3}) on frozen diffusion features; second, a Bi-GRU head trained with AdamW (1×10^{-4}, weight decay 1×10^{-5}) and label smoothing (0.1), again with the backbone frozen; and third, a fine-tuned setup in which early U-Net blocks remained frozen while deeper layers and a new spatial-temporal head were updated with differential learning rates (5e–5 for U-Net, 1e–3 for spatial extractor, 2e–3 for the GRU+MLP), including a 5-epoch warm-up and cosine learning-rate schedule. Finally, an ensemble late-fusion (0.1 radar : 0.9 RGB) combined one-hot logits from both branches.

IMPRESS NCSU [226] approach begins with the Frequency-Modulated Continuous Wave radar signals provided as RDM sequences. After describing each chirp $f_c(t) = f_0 + \frac{B}{T_c}t$ and mixing to obtain the IF signal $s_{IF}(t)$, they compute 2D FFTs across fast-time and slow-time to form RDMs. Rather than applying STFT to raw returns, they collapse the RDM along the range axis, summing bins $r_{\text{start}} = 50$ to $r_{\text{end}} = 200$, to derive equivalent micro-Doppler images $S_{\mu\text{-D}}(\omega, t)$. This yields a compact Doppler-time spectrogram per frame, which is then normalized and resized into 96×96 grayscale inputs. Classification is performed by a three-branch CNN: each branch processes one antenna's micro-Doppler image through two 64-filter and two 128-filter convolutional blocks (3×3 kernels, BatchNorm, ReLU, max-pool), then relational features are concatenated, passed through a 512-unit fully-connected layer with batch-norm, ReLU, and 0.5 dropout, and finally softmaxed over 126 classes. All models were trained using the Adam optimizer with an initial learning rate of 10^{-3} and standard cross-entropy loss. Inputs were batched at 64 samples per iteration, and training ran for 30 epochs with learning-rate reduction on plateau and early stopping based on validation accuracy. To ensure efficient use of the provided RDM data, the range-collapsed μ-D images were precomputed and cached, reducing on-the-fly FFT overhead. Model checkpoints were saved at each validation improvement, and the best weights were retained for final evaluation.

CPAMI (UW) [155] proposes FusionEnsemble-Net, a hierarchical ensemble that ingests both RGB video and RDM data through four parallel spatiotemporal backbones (3D ResNet-18, MC3-18, R(2+1)D-18, Swin-B). Each backbone extracts modality-specific features, which are then temporally modeled via LSTM or transformer encoders. An attention fusion module dynamically weights and combines the RGB and radar representations for each backbone, producing four fused feature vectors. Finally, each vector is classified by its own linear head, and the per-head probability outputs are averaged to yield the final sign prediction. The entire network was implemented in PyTorch and trained end-to-end on two NVIDIA A6000 GPUs. All spatiotemporal backbones were initialized with Kinetics-400 (for 3D CNNs) or ImageNet (for Swin-B) pre-trained weights. The RGB clips and the synchronized three-antenna RDM sequences were resized to 224×224, normalized, and batched

at size 4. They optimized the cross-entropy loss using AdamW with a cosine annealing learning-rate schedule, running for 30 epochs. Temporal modeling layers comprised two-layer LSTMs (hidden size 512) or transformer encoders (8 heads). During inference, predictions from all four classification heads were averaged, resulting in a robust ensemble output.

21.4 Discussion

This paper summarized *SignEval 2025*, the first multimodal SLR challenge organized within the MSLR workshop at ICCV 2025. The benchmark was structured around two complementary tracks that deliberately stress different generalization axes in realistic settings: a CSLR track based on pose-only inputs from the *Isharah* dataset, and an ISLR track based on the multimodal *MultiMeDaLIS* dataset. Within CSLR, the two shared tasks, *Signer-Independent* and *Unseen-Sentences*, isolated two of the most deployment-relevant requirements for continuous signing systems: robustness to signer variability and the ability to handle novel sentence compositions without relying on memorization of training sentences.

Beyond reporting leaderboards, the challenge design provided an explicit testbed for assessing which modeling choices translate into generalizable gains. The diversity of submitted solutions, spanning deep architectures, attention-based temporal modeling, and modality-specific adaptations, suggests that performance improvements are increasingly driven by how effectively systems encode temporal structure and exploit inductive biases aligned with signing dynamics. In the CSLR track, the results support the view that pose representations can be a viable foundation for scalable systems: by abstracting away appearance and background factors, pose-based pipelines can reduce overfitting to signer identity and recording conditions, while also lowering computational costs and offering a natural privacy advantage. At the same time, the benchmark also highlights open limitations of pose-only inputs, including sensitivity to pose extraction errors, partial occlusions, and the loss of fine-grained cues that may be needed for disambiguation in challenging sentence-level segments.

In the ISLR track, the outcomes reinforce the promise of multimodal sensing for robustness. *MultiMeDaLIS* enabled methods to move beyond vision-only assumptions by combining complementary modalities and exploring fusion strategies under constrained sensor availability. The strongest solutions indicate that sensor complementarity can mitigate failure modes of single-modality systems, which is especially relevant for medical communication contexts where privacy, environmental variability, and operational constraints can be binding. However, the benchmark also makes clear that fusion is not automatically beneficial: gains depend on principled alignment across modalities, handling missing or degraded streams, and evaluation protocols that expose robustness rather than only average accuracy.

Overall, *SignEval 2025* establishes a first shared reference point for multimodal SLR under standardized rules and highlights research directions that are difficult to study without a common

benchmark. The challenge framework supports inclusive and privacy-aware SLR by encouraging representations and sensing choices that can be deployed in sensitive environments, and by emphasizing generalization over dataset-specific optimization. At the same time, it motivates deeper investigation of (i) error sources inherited from upstream pose extraction, (ii) fairness and performance stratification across signers and linguistic constructions, and (iii) interaction-ready temporal stability beyond offline transcription metrics.

Looking forward, the MSLR challenge will continue with a second edition at CVPR 2026, with an expanded scope and updated resources. In particular, *MultiMeDaLIS* will be extended with acquisitions in multiple scenarios and sensor-to-signer configurations, and will include additional inertial streams from accelerometers and gyroscopes to support more robust multimodal modeling under visually challenging conditions. This next iteration will also facilitate new research themes such as continuous multimodal recognition, signer adaptation, and multilingual sign language understanding, further strengthening the benchmark as a community instrument for reproducible progress in SLR.

Chapter 22

Conclusion and Future Work

22.1 Cardiovascular Diagnostics: Multi-View Angiography and Clinical Trials

Part I investigated angiography-based assessment of coronary stenosis severity and functional significance FFR/iFR through spatio-temporal learning, generalization strategies, and deployment-oriented constraints. When reframed through the unifying lens of this dissertation, multimodality as an acquisition and robustness principle, the main conclusion is that progress toward routine clinical impact will depend less on marginal architectural sophistication and more on how effectively models can integrate complementary sources of evidence while preserving workflow feasibility and privacy constraints.

A first limitation of angiography-only inference is structural: a single projection is an inherently 2D and view-dependent representation of a 3D phenomenon. Overlap, foreshortening, and view-specific contrast patterns introduce ambiguity, particularly in intermediate lesions where clinical decision-making is most sensitive to uncertainty. A practical and clinically compatible direction is therefore multi-view integration whenever two or more projections are acquired. Multi-view fusion can be implemented in ways that remain cath-lab realistic: late fusion or cross-attention across view-specific encoders, coupled with view-quality gating to handle incomplete acquisitions and to down-weight unreliable projections. Beyond fusion, a stronger form of multimodality emerges when geometry is treated as an additional information channel: 3D-aware learning can enforce view consistency and introduce lightweight geometric priors through differentiable reconstruction components, epipolar constraints, or implicit 3D representations. Crucially, this step does not require new invasive measurements; it increases information content by leveraging what is often already available in routine practice, at the cost of careful protocol and robustness engineering.

A second conclusion is that FFR/iFR are not purely geometric quantities. Their physiological meaning depends on patient-specific hemodynamics, microvascular status, and procedural conditions

that are only partially observable from angiography. As a result, angiography-only models are expected to face irreducible uncertainty in a subset of cases. A multimodal pathway is therefore clinically motivated: keep angiography as the default input, but condition predictions on non-identifying physiological and procedural context when available and non-burdensome to collect. Even limited covariates (e.g., basic hemodynamic measurements, procedural metadata, acquisition settings) can reduce ambiguity in borderline cases, improve calibration, and support outputs that are interpretable in terms of confidence and decision thresholds. Where feasible, the longer-term opportunity is selective fusion with complementary signals that add functional information without disrupting workflow, with the explicit goal of improving reliability rather than maximizing raw accuracy.

Clinical translation ultimately requires evidence that reflects real decision processes. Retrospective benchmarks remain necessary but insufficient; the next steps with the highest translational value are external validation and prospective evaluation. Multi-center studies with held-out sites, heterogeneous protocols, and diverse equipment are required to quantify generalization and to surface failure modes under realistic acquisition variability. Prospective trials should then evaluate not only predictive performance but also operational impact: workflow integration, time overhead, consistency of decisions in intermediate lesions, and downstream outcomes. A particularly important component is human-AI interaction. Reader studies can test whether cardiologists meaningfully use uncertainty information, whether the system reduces inter-operator variability, and how model outputs interact with time pressure and institutional practice patterns. It should also be noted that this dissertation does not empirically evaluate human-AI interaction in the catheterization laboratory or the integration of the proposed models into routine interventional workflows. These aspects require dedicated prospective studies, including reader studies with interventional cardiologists, workflow simulations under time constraints, and analyses of how uncertainty-aware outputs influence clinical decisions. Therefore, clinical integration is treated here as a future translational step rather than as an experimentally validated outcome of the thesis.

For procedure support, performance must be coupled with calibration and uncertainty management as explicit safety properties. Future systems should provide well-calibrated probabilities, uncertainty estimates that are meaningful under domain shift, and conservative policies for abstention or referral in low-confidence regimes. These mechanisms should be quality-aware by design: acquisition quality factors (contrast opacification, motion, vessel overlap, projection suitability) should be estimated explicitly and propagated to decision thresholds, limiting overconfident predictions when the input does not support reliable inference. In this dissertation's multimodal framing, quality indicators are not ancillary metadata; they are part of the input evidence that determines whether a prediction should be trusted.

Finally, scaling to the diversity required for robust cardiology AI remains difficult without distributed collaboration. Federated and privacy-preserving learning provides a natural route to aggregate knowledge across institutions while respecting governance constraints. However, non-

IID heterogeneity in clinical data makes personalization and site-aware strategies important, and privacy mechanisms (secure aggregation, differential privacy where required) should be paired with evaluation protocols that preserve clinical meaning across sites. In particular, federated calibration and threshold selection deserve attention, because consistent operating points are central to clinical adoption and cannot be treated as an afterthought.

Overall, the direction suggested by Part I is to treat multimodality as the main driver of clinical robustness: integrate multiple angiographic views when available, incorporate lightweight physiological and procedural context to resolve functional ambiguity, and validate models prospectively with calibration-oriented and safety-oriented protocols, while preserving deployment realism and privacy constraints.

22.2 Cognitive and Behavioral Monitoring: Multimodal Fusion and Real-Time Deployment

Part II investigated automated cognitive and behavioral monitoring from scalable, non-invasive signals, spanning vision-derived behavioral cues (facial micro-dynamics and eye-gaze traces) and long-term wearable time series acquired in unconstrained daily life. Viewed through the dissertation's multimodal perspective, the main conclusion is that practical cognitive monitoring is achievable when sensing is low-burden and widely available, but translation requires progressing from single-modality, offline score estimation toward multimodal, quality-aware, and longitudinal decision support that remains reliable under heterogeneous acquisition and across diverse populations.

A first direction is multimodal behavioral fusion. Human behavior during cognitive tasks is intrinsically multidimensional: attention, engagement, affect, and motor control interact, and no single channel is expected to remain stable across subjects and environments. Future systems should integrate complementary cues into a unified temporal representation, combining facial dynamics, gaze trajectories, and head motion to characterize task engagement and response patterns. Wearable streams can provide additional, low-burden evidence through inertial and cardiac-related signals that capture rhythms, activity, and physiological state in everyday settings. Where consented and clinically appropriate, speech and language can contribute further information through prosody, fluency, and lexical markers, while interaction logs (response times, errors, navigation patterns, task completion signals) provide structured features that are often less privacy-sensitive than raw audio-video. From an engineering standpoint, the key requirement is graceful operation under missing modalities: models should degrade predictably when a channel is unavailable or unreliable, and remain deployable across devices without strict hardware assumptions.

A second challenge concerns robustness and equity under domain shift. Vision-based monitoring is sensitive to camera quality, illumination, viewpoint, occlusions, and compression; wearable signals are affected by device differences, placement variability, and irregular sampling. At the same time,

behavioral signatures vary with age, culture, education, and clinical phenotype. Consequently, the most meaningful advances will come from stronger evidence of generalization rather than incremental gains on a single dataset. This is particularly important because the cognitive monitoring experiments presented in this dissertation rely on limited and project-specific datasets, which support methodological feasibility but do not yet provide definitive evidence of generalization across sites, devices, and populations. Future work should prioritize cross-device and cross-setting evaluation, together with population-aware protocols that quantify performance differences across demographic and clinical subgroups. For deployment, calibration transfer is central: models must preserve reliable confidence under changing acquisition conditions and avoid overconfident outputs when the observed distribution departs from training. This is not only a technical requirement but also an equity requirement, because systematic performance gaps can propagate into unequal screening and referral pathways.

Beyond single-session inference, cognitive decline and behavioral change are longitudinal by nature. Clinically useful tools should therefore treat time as a first-class dimension, shifting from population-level boundaries to subject-aware baselines that model each individual's typical behavioral signature and detect meaningful deviations. Trajectory modeling across repeated sessions can estimate trends and progression rates while representing uncertainty explicitly, which is essential when measurements are noisy and contexts vary. This perspective also motivates personalized monitoring for rare events in long-term data, where anomaly-style formulations and subject-specific normality models can better reflect inter-subject heterogeneity than global classifiers. Longitudinal deployment naturally raises the need for continual or incremental adaptation under privacy constraints, enabling systems to track evolving patterns without catastrophic forgetting, while retaining conservative behavior when uncertainty is high.

Translation also depends on deployability beyond offline accuracy. Interactive assessments and at-home monitoring require streaming-capable pipelines with low latency, stable decisions, and robustness to transient noise. On-device processing becomes a dual requirement: it reduces latency and improves reliability under connectivity constraints, while limiting transmission of identifiable signals and thereby strengthening privacy by design. A key enabling component is quality assessment: systems should detect low-quality acquisition conditions (poor lighting, unstable tracking, excessive motion, partial occlusion, unreliable wearable contact) and respond conservatively by abstaining, requesting re-acquisition, or switching to signals that remain reliable. In this setting, confidence estimation and safe abstention are not auxiliary features but core safety mechanisms for tools that may influence screening and referral decisions.

Impact ultimately depends on clinically grounded validation and human-centered design. Prospective studies should compare AI outputs against clinician-administered assessments over time, emphasizing repeated-measure consistency, clinically meaningful thresholds, and decision-relevant endpoints such as screening triggers and referral criteria, rather than only aggregate correlations. Interfaces should support trust without overstating certainty: instead of exposing raw scores alone,

systems should provide actionable summaries, uncertainty-aware explanations, and feedback that is transparent about limitations while minimizing unnecessary anxiety for users. Although this dissertation discusses human-centered deployment as a necessary condition for cognitive monitoring, the presented studies do not yet include empirical user studies with patients, caregivers, or clinicians, nor do they evaluate integration into real screening workflows. Future work should therefore include usability studies, clinician-in-the-loop validation, and longitudinal deployment protocols to assess whether the proposed outputs are understandable, actionable, and beneficial in realistic assessment pathways. Without this clinical and human-centered framing, even strong multimodal models risk remaining confined to controlled demonstrations.

Overall, Part II motivates a roadmap centered on multimodal fusion robust to missing channels, fairness-aware generalization across devices and populations, longitudinal subject-aware modeling, edge-friendly privacy-aware deployment, and prospective validation aligned with real clinical decisions.

22.3 Sign-Language Recognition: Real-Time Clinical Integration and User Studies

Part III addressed sign language recognition for medical communication through multimodal dataset design, privacy-preserving sensing with radar, semantic grounding via text alignment, and community benchmarking. Taken together, these contributions support the feasibility of assistive SLR under realistic acquisition constraints where privacy is a binding requirement rather than an afterthought. At the same time, clinical translation depends on properties that are only partially captured by offline benchmarks: interaction-ready inference, robustness to signer and environment variability, clinically meaningful outputs, and human-centered validation aligned with how communication unfolds in healthcare settings.

A first requirement for clinical relevance is the shift from offline classification to real-time interaction. Patient-doctor exchanges demand streaming inference with bounded latency and stable temporal behavior, particularly under continuous signing and conversational turn-taking. Future systems should therefore support streaming recognition with principled temporal smoothing to reduce jitter, and incremental decoding that updates hypotheses as evidence accumulates, enabling timely and natural system responses. From a deployment standpoint, on-device/edge inference is equally important because it reduces privacy exposure associated with transmitting identifiable streams and lowers infrastructure requirements in hospitals. Real-time operation must explicitly handle interruptions, partial signs, and variable pacing under stress, which are common in clinical encounters and can induce failure modes that remain hidden in curated offline data.

A key future direction is the transition from isolated medical signs to continuous clinical dialogue. This step cannot be achieved by simply applying an isolated classifier frame by frame, because

conversational signing includes coarticulation, implicit segmentation boundaries, signer-dependent rhythm, and non-manual grammatical markers that interact with manual signs. Real-time clinical translation will therefore require streaming models that jointly estimate temporal boundaries, recognize manual and non-manual components, and maintain dialogue context for disambiguation. In the medical setting, this is especially important for negation, pain intensity, temporal references, and symptom descriptions, where small linguistic differences can change the clinical meaning. The contributions in this dissertation provide dataset, sensing, and representation foundations for privacy-aware medical SLR, but full conversational translation remains a broader research objective requiring larger continuous corpora, linguistically informed annotation, and user studies with Deaf signers, interpreters, and clinicians.

Robust deployment also hinges on generalization across signers, environments, and domain vocabulary. Variability in signing style, demographic and linguistic factors, camera setups, and clinical context can markedly affect performance. Future work should prioritize cross-signer protocols with fairness analysis, while expanding from isolated medical vocabulary toward richer patient-doctor dialogues where meaning is context-dependent. Robustness also requires explicit attention to occlusions and constrained viewpoints typical of clinical spaces (e.g., bedside interactions, crowded environments), together with scenario-driven benchmarking across sensor placements and acquisition configurations to quantify sensitivity to viewpoint and environment shifts. In this direction, releasing an extended dataset version (*MultiMeDaLIS 2*) acquired across multiple scenarios and sensor-to-signer configurations, and enriched with synchronized wrist/hand inertial streams (accelerometer and gyroscope), is a pragmatic step: it enables controlled robustness studies and provides complementary motion evidence that can improve discrimination when visual conditions are suboptimal.

A distinctive dimension of Part III is the focus on privacy-preserving sensing and alternative modalities. Radar-based acquisition provides a strong privacy advantage by reducing identifiable visual content, but it also introduces limitations, including reduced spatial detail compared to RGB and sensitivity to environmental factors. Future work can strengthen this direction by improving radar representations through learned preprocessing and self-supervised objectives that better exploit range-angle and Doppler dynamics, and by advancing multimodal privacy-aware fusion that combines radar with less identifiable auxiliary signals such as inertial sensing. Robustness to multipath reflections, sensor placement changes, and interference remains particularly important in clinical environments, where room geometry and equipment can vary widely. In parallel, translation requires edge-efficient modeling: real-time radar pipelines must operate under tight memory and power budgets. A concrete line of work is to adapt data-free post-training quantization methods such as our HC-PTQ approach [243] to radar-based SLR backbones, targeting dependable edge deployment while preserving recognition performance.

In healthcare applications, recognition accuracy alone is rarely sufficient: systems must also produce clinically meaningful outputs. Predicting isolated labels may not capture communicative intent, ambiguity, or the cost of misunderstandings. Future systems should map recognized signing

to structured intents or standardized phrases, support context-aware disambiguation using dialogue history and clinician prompts (and, when appropriate, limited patient metadata), and expose uncertainty-aware outputs that request clarification when ambiguity is high. Text alignment and vision-language objectives provide a promising foundation for semantic grounding, but clinical use demands interface design and evaluation protocols that explicitly reflect the risk of miscommunication at the operating points where the tool is intended to be used.

Real-world integration should be guided by human-centered evaluation. Co-design with Deaf users, clinicians, and interpreters is needed to define realistic use cases, acceptable failure modes, and interaction patterns that are usable under time pressure. User studies in clinical-like environments should measure usability, trust, cognitive load, and perceived privacy. Outcome-oriented evaluation should then quantify whether the system reduces communication barriers and improves task completion in realistic pathways, rather than relying solely on aggregate recognition metrics. At the same time, the studies presented in this dissertation do not yet empirically evaluate the integration of SLR systems into real patient-doctor interactions. Future work should therefore include co-design sessions and user studies with Deaf signers, clinicians, and interpreters, assessing not only recognition accuracy but also communication effectiveness, latency tolerance, perceived privacy, trust, and the clinical consequences of recognition errors.

Sustained progress in assistive SLR also depends on reproducible community benchmarking. Shared protocols, standardized splits, and openly comparable baselines are essential to separate methodological improvements from dataset-specific effects. Community initiatives such as workshops and challenges provide a structured mechanism to stress-test generalization under controlled rules and to consolidate evaluation practices. As a next step, a workshop (2nd MSLR, Multimodal Sign Language Recognition) and challenge (2nd SignEval) based on the frequency modalities of *MultiMeDaLIS* [248] dataset have been accepted at the IEEE/CVF CVPR 2026 [135], building on the first edition at IEEE/CVF ICCV 2025 [222] and aiming to consolidate evaluation practices for privacy-aware multimodal sign recognition in realistic settings.[1]

Overall, Part III reinforces the dissertation's acquisition-aware multimodal thesis: preserving the clinical objective of accessible communication while widening the feasible pathway through privacy-preserving sensing, multimodal fusion, edge-efficient modeling, and clinically grounded interaction and user validation.

[1] https://multimodal-sign-language-recognition.github.io/CVPR-2026/

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Accountability Act et al. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[3] Oladipupo O Adeoluwa, Sean J Kearney, Emre Kurtoglu, Charles J Connors, and Sevgi Z Gurbuz. Near real-time asl recognition using a millimeter wave radar. In *Radar Sensor Technology XXV*, volume 11742, pages 173–184. SPIE, 2021.

[4] Yousef Alimohamadi, Mojtaba Sepandi, Maryam Taghdir, and Hadiseh Hosamirudsari. Determine the most common clinical symptoms in covid-19 patients: a systematic review and meta-analysis. *Journal of preventive medicine and hygiene*, 61(3):E304, 2020.

[5] Sarah Alyami and Hamzah Luqman. Swin-mstp: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617:129015, 2025.

[6] Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. Isharah: A large-scale multi-scene dataset for continuous sign language recognition. *arXiv preprint arXiv:2506.03615*, 2025.

[7] Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects. *Information Processing & Management*, 61(5):103774, 2024.

[8] Tim J Anderson and Michael R MacAskill. Eye movements in patients with neurodegenerative disorders. *Nature Reviews Neurology*, 9(2):74–85, 2013.

[9] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[10] Karol Antczak and Łukasz Liberadzki. Stenosis detection with deep convolutional neural networks. In *MATEC web of conferences*, volume 210, page 04001. EDP Sciences, 2018.

[11] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020.

[12] Homa Arab, Iman Ghaffari, Lydia Chioukh, Serioja Ovidiu Tatu, and Steven Dufour. A convolutional neural network for human motion recognition and classification using a millimeter-wave doppler radar. *IEEE Sensors Journal*, 22(5):4494–4502, 2022.

[13] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[14] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[15] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[16] Benjamin Au, Uri Shaham, Sanket Dhruva, Georgios Bouras, Ecaterina Cristea, Andreas Coppi, Fred Warner, Shu-Xia Li, and Harlan Krumholz. Automated characterization of stenosis in invasive coronary angiography images with convolutional neural networks. *arXiv preprint arXiv:1807.10597*, 2018.

[17] Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, page 103000, 2023.

[18] J. L. Ba, J. R. Kiros, and G. E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[19] Shaojie Bai. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[20] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[21] A. Bandini, S. Orlandi, H.J. Escalante, F. Giovannelli, M. Cincotta, C.A. Reyes-Garcia, P. Vanni, G. Zaccara, and C. Manfredi. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *Journal of neuroscience methods*, 281:7–20, 2017.

[22] Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. Non-invasive classification of alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR, 2020.

[23] Stefan Baumann, Leonard Chandra, Elizaveta Skarga, Matthias Renker, Martin Borggrefe, Ibrahim Akin, and Dirk Lossnitzer. Instantaneous wave-free ratio (ifr®) to determine hemodynamically significant coronary stenosis: a comprehensive review. *World Journal of Cardiology*, 2018.

[24] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[25] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Shadi Albarqouni. Feddis: Disentangled federated learning for unsupervised brain pathology segmentation. *arXiv preprint arXiv:2103.03705*, 2021.

[26] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38 th International Conference on Machine Learning*, volume 1, 2021.

[27] Carmen Bisogni, Michele Nappi, Genoveffa Tortora, and Alberto Del Bimbo. Gaze analysis: A survey on its applications. *Image and Vision Computing*, 144:104961, 2024.

[28] Sabyasachi Biswas, Ahmed Manavi Alam, and Ali C Gurbuz. Hrspecnet: A deep learning-based high-resolution radar micro-doppler signature reconstruction for improved har classification. *IEEE Transactions on Radar Systems*, 2:484–497, 2024.

[29] Sabyasachi Biswas, Cemre Omer Ayna, and Ali C Gurbuz. Plfnets: Interpretable complex valued parameterized learnable filters for computationally efficient rf classification. *IEEE Transactions on Radar Systems*, 2024.

[30] J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[31] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

[32] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[33] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191, 2022.

[34] Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[35] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[36] Chiara Branchini, Lara Mantovan, et al. *A Grammar of Italian Sign Language (LIS)*. Edizioni Ca'Foscari, 2020.

[37] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[38] Francesco Bruno, Fabrizio D'Ascenzo, Giorgio Marengo, Roberto Manfredi, Andrea Saglietto, Guglielmo Gallone, Luca Franchin, Francesco Piroli, Filippo Angelini, Ovidio De Filippo, et al. Fractional flow reserve guided versus angiographic guided surgical revascularization: A meta-analysis. *Catheterization and Cardiovascular Interventions*, 98(1):E18–E23, 2021.

[39] Salvatore Calcagno, Raffaele Mineo, Daniela Giordano, and Concetto Spampinato. Ensemble and personalized transformer models for subject identification and relapse detection in e-prevention challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023.

[40] Gaia Caligiore. *Codifying the body: exploring the cognitive and socio-semiotic framework in building a multimodal Italian sign language (LIS) dataset*. PhD thesis, University of Catania, 2024.

[41] Gaia Caligiore, Raffaele Mineo, Concetto Spampinato, Egidio Ragonese, Simone Palazzo, and Sabina Fontana. Multisource approaches to italian sign language (LIS) recognition: Insights from the multimedalis dataset. In Felice DellOrletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 132–140, Pisa, Italy, December 2024. CEUR Workshop Proceedings.

[42] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.

[43] Paolo G Camici, Giulia d'Amati, and Ornella Rimoldi. Coronary microvascular dysfunction: mechanisms and functional assessment. *Nature Reviews Cardiology*, 12(1):48–62, 2015.

[44] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[45] Anna Cardinaletti and Lara Mantovan. Le lingue dei segni nel "volume complementare" e l'insegnamento della lis nelle università italiane. *Italiano LinguaDue*, 14(2):113–128, 2022.

[46] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[47] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[48] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[49] Marco Cascella, Michael Rajnik, Abdul Aleem, Scott C Dulebohn, and Raffaela Di Napoli. Features, evaluation, and treatment of coronavirus (covid-19). *Treasure Island (FL): StatPearls Publishing*, 2020.

[50] Carlo Cecchetto, Serena Giudice, Emiliano Mereghetti, et al. La raccolta del corpus lis. In *Grammatica, lessico e dimensioni di variazione nella LIS*, pages 55–67. Franco Angeli, 2011.

[51] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[52] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[53] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[54] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.

[55] Tianyang Chen, Xichao Dong, and Yaowen Chen. Gesture recognition with feature fusion using fmcw radar. In *Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021)*, volume 12166, pages 353–362. SPIE, 2022.

[56] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[57] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 697–714. Springer, 2020.

[58] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[59] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[60] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[61] Adriaan Coenen, Young-Hak Kim, Mariusz Kruk, Christian Tesche, Jakob De Geer, Akira Kurata, Marisa L Lubbers, Joost Daemen, Lucian Itu, Saikiran Rapaka, et al. Diagnostic accuracy of a machine-learning approach to coronary computed tomographic angiography–based fractional flow reserve: result from the machine consortium. *Circulation: Cardiovascular Imaging*, 11(6):e007217, 2018.

[62] Asher Cohen, John A Naslund, Sarah Chang, Srilakshmi Nagendra, Anant Bhan, Abhijit Rozatkar, Jagadisha Thirthalli, Ameya Bondre, Deepak Tugnawat, Preethi V Reddy, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during covid-19: a prospective, three-site, two-country, longitudinal study. *Schizophrenia*, 9(1):6, 2023.

[63] C. Collet, J. Sonck, B. Vandeloo, T. Mizukami, and B. et al. Roosens. Measurement of Hyperemic Pullback Pressure Gradients to Characterize Patterns of Coronary Atherosclerosis. *J Am Coll Cardiol*, 74, 2019.

[64] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery*, 102(3):148–158, 2015.

[65] Chao Cong, Yoko Kato, Henrique D Vasconcellos, Mohammad R Ostovaneh, Joao AC Lima, and Bharath Ambale-Venkatesh. Deep learning-based end-to-end automated stenosis classification and localization on catheter coronary angiography. *medRxiv*, 2021.

[66] Chao Cong, Yoko Kato, Henrique Doria Vasconcellos, Joao Lima, and Bharath Venkatesh. Automated stenosis detection and classification in x-ray angiography using deep neural network. In *2019 IEEE BIBM*, 2019.

[67] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7361–7369, 2017.

[68] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

[69] Viacheslav Danilov, Olga Gerget, Kirill Klyshnikov, Evgeny Ovcharenko, and Alejandro Frangi. Comparative study of deep learning models for automatic coronary stenosis detection in x-ray angiography. In *Proceedings of the 30th International Conference on Computer Graphics and Machine Vision*, 2020.

[70] Viacheslav V Danilov, Kirill Yu Klyshnikov, Olga M Gerget, Anton G Kutikhin, Vladimir I Ganyukov, Alejandro F Frangi, and Evgeny A Ovcharenko. Real-time coronary artery stenosis detection based on modern neural networks. *Scientific reports*, 11(1):1–13, 2021.

[71] Justin E Davies, Sayan Sen, Hakim-Moulay Dehbi, Rasha Al-Lamee, Ricardo Petraco, Sukhjinder S Nijjer, Ravinay Bhindi, Sam J Lehman, Darren Walters, James Sapontis, et al. Use of the instantaneous wave-free ratio or fractional flow reserve in pci. *New England Journal of Medicine*, 376(19):1824–1834, 2017.

[72] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[73] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.

[74] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-temporal graph convolutional networks for sign language recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer, 2019.

[75] Bernard De Bruyne, Nico HJ Pijls, Bindu Kalesan, Emanuele Barbato, Pim AL Tonino, Zsolt Piroth, Nikola Jagic, Sven Möbius-Winkler, Gilles Rioufol, Nils Witt, et al. Fractional flow

reserve–guided pci versus medical therapy in stable coronary disease. *New England Journal of Medicine*, 367(11):991–1001, 2012.

[76] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450, 2021.

[77] Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Angiographic image processing method and system, 2025. WO Patent 2025262634A1.

[78] Gaetano Maria De Ferrari, Fabrizio D'Ascenzo, Ovidio De Filippo, Mario Iannaccone, Marco Aldinucci, James Michael Hughes, Michele Millesimo, Walter Grosso Marra, Giorgio Quadri, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, Raffaele Mineo, and Simone Palazzo. Procedimento e sistema di elaborazione di immagini angiografiche, 2026. IT Patent 202400014434A1.

[79] Ovidio De Filippo, Raffaele Mineo, Michele Millesimo, Wojciech Wańha, Federica Proietto Salanitri, Antonio Greco, Antonio Maria Leone, Luca Franchin, Simone Palazzo, Giorgio Quadri, et al. Non-invasive physiological assessment of intermediate coronary stenoses from plain angiography through artificial intelligence: the starflow system. *European Heart Journal-Quality of Care and Clinical Outcomes*, 11(3):343–352, 2025.

[80] Tullio De Mauro and Isabella Chiari. Il nuovo vocabolario di base della lingua italiana. *Internazionale. it*, 2016.

[81] Bidya Debnath, Iffat Ara Ebu, Sabyasachi Biswas, Ali C Gurbuz, and John E Ball. Fmcw radar range profile and micro-doppler signature fusion for improved traffic signaling motion classification. In *2024 IEEE Radar Conference (RadarConf24)*, pages 1–6. IEEE, 2024.

[82] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[83] R. Dias and A. Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, 11, 2019.

[84] Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer. *Serious games*. Springer, 2016.

[85] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[86] NR Draper. Applied regression analysis, 1998.

[87] Tianming Du, Xuqing Liu, Honggang Zhang, and Bo Xu. Real-time lesion detection of cardiac coronary artery using deep neural networks. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 150–154. IEEE, 2018.

[88] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[89] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[90] Paul Ekman. Lie catching and microexpressions. *The philosophy of deception*, 1(2):5, 2009.

[91] El-Sayed M El-Alfy and Hamzah Luqman. A comprehensive survey and taxonomy of sign language research. *Engineering Applications of Artificial Intelligence*, 114:105198, 2022.

[92] Fatimah Mohamed Emad Elden. Cslrconformer: A data-centric conformer approach for continuous arabic sign language recognition on the isharah datase. *arXiv preprint arXiv:2508.01791*, 2025.

[93] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[94] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[95] Biyi Fang, Jillian Co, and Mi Zhang. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM conference on embedded network sensor systems*, pages 1–13, 2017.

[96] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[97] Robin Faro, Francesco Cancelliere, Alesssandro Strano, and Raffaele Mineo. Timely predicting power mosfet failure in ev: Ai based metodology and deployment at the edge. In *2025 IEEE Symposia on Computational Intelligence for Energy, Transport and Environmental Sustainability (CIETES Companion)*, pages 1–5. IEEE, 2025.

[98] Hassan Ismail Fawaz. Deep learning for time series classification. *arXiv preprint arXiv:2010.00567*, 2020.

[99] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[100] W. F Fearon, B. Bornschein, P. AL Tonino, R. M Gothe, B. De Bruyne, N. HJ Pijls, and U. Siebert. Economic evaluation of fractional flow reserve–guided percutaneous coronary intervention in patients with multivessel disease. *Circulation*, 122, 2010.

[101] William F Fearon, Takeshi Nishi, Bernard De Bruyne, Derek B Boothroyd, Emanuele Barbato, Pim Tonino, Peter Jüni, Nico HJ Pijls, and Mark A Hlatky. Clinical outcomes and cost-effectiveness of fractional flow reserve–guided percutaneous coronary intervention in patients with stable coronary artery disease: three-year follow-up of the fame 2 trial (fractional flow reserve versus angiography for multivessel evaluation). *Circulation*, 137(5):480–487, 2018.

[102] Ines Feki, Sourour Ammar, Yousri Kessentini, and Khan Muhammad. Federated learning for covid-19 screening from chest x-ray images. *Applied Soft Computing*, 106:107330, 2021.

[103] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *CVPR*, 2018.

[104] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

[105] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

[106] Sabina Fontana and Gaia Caligiore. Italian sign language (lis) and natural language processing: an overview. *NL4AI@ AI* IA*, 2021.

[107] Sabina Fontana, Serena Corazza, Penny Boyes Braem, and Virginia Volterra. Language research and language community change: Italian sign language 1981–2013. *International Journal of the Sociology of Language*, 2015(236):1–30, 2015.

[108] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916, 2014.

[109] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8025–8035, 2024.

[110] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[111] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

[112] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178, 2009.

[113] Carlo Geraci, Katia Battaglia, Anna Cardinaletti, Carlo Cecchetto, Caterina Donati, Serena Giudice, and Emiliano Mereghetti. The lis corpus project: A discussion of sociolinguistic variation in the lexicon. *Sign Language Studies*, 11(4):528–574, 2011.

[114] J.O. Gerlowska, K. Dmitruk, and K. Rejdak. Facial emotion mimicry in older adults with and without cognitive impairments due to alzheimer's disease. *AIMS Neurosci*, 2017.

[115] Gabriele Gianfreda et al. Un corpus di conversazioni in lingua dei segni italiana attraverso videochat: una proposta per la loro trascrizione e analisi. In *Grammatica, lessico e dimensioni di variazione nella LIS*, pages 95–109. FrancoAngeli, 2011.

[116] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*, pages 201–210. PMLR, 2016.

[117] W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[118] Matthias Götberg, Evald H Christiansen, Ingibjörg Gudmundsdottir, Lennart Sandhall, Elmir Omerovic, Stefan K James, David Erlinge, and Ole Fröbert. Instantaneous wave-free ratio versus fractional flow reserve guided intervention (ifr-swedeheart): rationale and design of a multicenter, prospective, registry-based randomized clinical trial. *American heart journal*, 170(5):945–950, 2015.

[119] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay. M3d-cam: A pytorch library to generate 3d data attention maps for medical deep learning, 2020.

[120] M. J Grundeken, C. Collet, Y. Ishibashi, P. Généreux, T. Muramatsu, and L. et al. LaSalle. Visual estimation versus different quantitative coronary angiography methods to assess lesion severity in bifurcation lesions. *Catheterization and Cardiovascular Interventions*, 91, 2018.

[121] M. J. Grundeken, C. Collet, Y. Ishibashi, P. reux, and T. et al. Muramatsu. Visual estimation versus different quantitative coronary angiography methods to assess lesion severity in bifurcation lesions. *Catheter Cardiovasc Interv*, 91, 2018.

[122] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros,

et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[123] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[124] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[125] Sevgi Z Gurbuz, Ali C Gurbuz, Evie A Malaia, Darrin J Griffin, Chris Crawford, M Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafi, Ajaymehul Anbuselvam, et al. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE international radar conference (RADAR)*, 2020.

[126] Sevgi Zubeyde Gurbuz and Moeness G Amin. Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring. *IEEE Signal Processing Magazine*, 36(4):16–28, 2019.

[127] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[128] J. Hamm, C.G. Kohler, R.C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, 200(2):237–256, 2011.

[129] T. Han, D. Ai, X. Li, and J. et al. Fan. Coronary artery stenosis detection via proposal-shifted spatial-temporal transformer in x-ray angiography. *Computers in Biology and Medicine*, 2023.

[130] Tao Han, Danni Ai, Xinyu Li, Jingfan Fan, Hong Song, Yining Wang, and Jian Yang. Coronary artery stenosis detection via proposal-shifted spatial-temporal transformer in x-ray angiography. *Computers in Biology and Medicine*, 2023.

[131] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[132] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

[133] Md Rezwanul Haque, Md. Milon Islam, S M Taslim Uddin Raju, and Fakhri Karray. A signer-invariant conformer and multi-scale fusion transformer for continuous sign language

recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[134] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[135] Ahmed Abul Hasanaath, Raffaele Mineo, Hamzah Luqman, Sarah Alyami, Maad Alowaifeer, Amelia Sorrenti, Gaia Caligiore, Sabina Fontana, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, Motaz Alfarraj, Mufti Mahmud, Simone Palazzo, and Nour Imane Zeghib. Signeval 2026 challenges results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026.

[136] Saad Hassan, Larwan Berke, Elahe Vahdani, Longlong Jing, Yingli Tian, and Matt Huenerfauth. An isolated-signing rgbd dataset of 100 american sign language signs produced by fluent asl signers. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 89–94, 2020.

[137] H He and EA Garcia. Learning from imbalanced data ieee transactions on knowledge and data engineering, 2009.

[138] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

[139] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[140] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[141] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[142] Stephen L Hicks, Rakesh Sharma, Amad N Khan, Claire M Berna, Andrea Waldecker, Kevin Talbot, Chris Kennard, and Martin R Turner. An eye-tracking version of the trail-making test. *PloS one*, 8(12):e84061, 2013.

[143] Marlene Hilzensauer and Klaudia Krammer. A multilingual dictionary for sign languages:" spreadthesign". In *ICERI2015 Proceedings*, pages 7826–7834. IATED, 2015.

[144] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[145] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[146] Ashley A Holmes, Shikha Tripathi, Emily Katz, Ijah Mondesire-Crump, Rahul Mahajan, Aaron Ritter, Teresa Arroyo-Gallego, and Luca Giancardo. A novel framework to estimate cognitive impairment via finger interaction with digital devices. *Brain Communications*, 4(4):fcac194, 2022.

[147] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[148] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[149] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239, 2023.

[150] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539, 2023.

[151] Lisa I Iezzoni, Bonnie L O'Day, Mary Killeen, and Heather Harker. Communicating about health care: observations from persons who are deaf or hard of hearing. *Annals of internal medicine*, 140(5):356–362, 2004.

[152] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[153] Infineon Technologies. Infineon® BGT60TR13C XENSIV™ 60GHz radar sensor for advanced sensing.

[154] Intel Corporation. Intel® RealSense™ Depth Camera D455: Twice the Range. Better Performance. IMU. Global Shutter RGB. Up to 4 m with Z-Error ¡ 2%. On-chip calibration.

[155] Md. Milon Islam, Md Rezwanul Haque, S M Taslim Uddin Raju, and Fakhri Karray. Fusionensemble-net: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[156] Lucian Itu, Saikiran Rapaka, Tiziano Passerini, Bogdan Georgescu, Chris Schwemmer, Max Schoebinger, Thomas Flohr, Puneet Sharma, and Dorin Comaniciu. A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. *Journal of applied physiology*, 121(1):42–52, 2016.

[157] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[158] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[159] Deepthi Jallepalli, Navya Chennagiri Ravikumar, Poojitha Vurtur Badarinath, Shravya Uchil, and Mahima Agumbe Suresh. Federated learning for object detection in autonomous vehicles. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 107–114. IEEE, 2021.

[160] Hyeju Jang, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Saffrin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, Pavan Tutt, et al. Classification of alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-movement data. *Frontiers in Human Neuroscience*, 15:716670, 2021.

[161] Yu-Chiao Jhaung, Yu-Ming Lin, Chiao Zha, Jenq-Shiou Leu, and Mario Köppen. Implementing a hand gesture recognition system based on range-doppler map. *Sensors*, 22(11):4260, 2022.

[162] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[163] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.

[164] C. Jiang, J. Wu, W. Zhong, M. Wei, J. Tong, H. Yu, and L. Wang. Automatic facial paralysis assessment via computational image analysis. *Journal of Healthcare Engineering*, 2020.

[165] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423, 2021.

[166] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[167] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20676–20686, 2023.

[168] Longlong Jing, Elahe Vahdani, Matt Huenerfauth, and Yingli Tian. Recognizing american sign language manual signs from rgb-d videos. *arXiv preprint arXiv:1906.02851*, 2019.

[169] Samuel Johnny, Blessed Guda, Andrew Stephen, and Assane Gueye. Autosign: Direct pose-to-text translation for continuous sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[170] Nils P Johnson, Daniel T Johnson, Richard L Kirkeeide, Colin Berry, Bernard De Bruyne, William F Fearon, Keith G Oldroyd, Nico HJ Pijls, and K Lance Gould. Repeatability of fractional flow reserve despite variations in systemic and coronary hemodynamics. *JACC: Cardiovascular Interventions*, 8(8):1018–1027, 2015.

[171] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[172] Jakub Juranek. Multimodal italian sign language recognition with radar–video late fusion on the multimedalis dataset. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[173] C Jyotsna, Joseph Amudha, Amritanshu Ram, Damiano Fruet, and Giandomenico Nollo. Predicteye: Personalized time series model for mental state prediction using eye tracking. *IEEE Access*, 11:128383–128409, 2023.

[174] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[175] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[176] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[177] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[178] S M Kazemi, R Goel, S Eghbali, J Ramanan, J Sahota, S Thakur, S Wu, C Smyth, P Poupart, and M Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

[179] Zhenghao Ke, Sheng Liu, and Yuan Feng. Fine-grained cross-modality consistency mining for continuous sign language recognition. *Pattern Recognition Letters*, 191:23–30, 2025.

[180] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.

[181] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE transactions on biomedical engineering*, 63(3):664–675, 2015.

[182] Sivan Klil-Drori, Katie C Bodenstein, Shuo Sun, Lara Kojok, Johanna Gruber, Youssef Ghantous, Jeffrey Cummings, and Ziad Nasreddine. Montreal cognitive assessment (moca) xpresso: Validation of a digital self-administered cognitive prescreening tool. *Journal of the American Geriatrics Society*, 2024.

[183] J Knuuti, W Wijns, A Saraste, D Capodanno, E Barbato, C Funck-Brentano, et al. Esc guidelines for the diagnosis and management of chronic coronary syndromes [published correction appears in eur heart j. 2020 nov 21; 41 (44): 4242]. *EHJ*, 41, 2020.

[184] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[185] Bon-Kwon Koo, Andrejs Erglis, Joon-Hyung Doh, David V Daniels, Sanda Jegere, Hyo-Soo Kim, Allison Dunning, Tony DeFrance, Alexandra Lansky, Jonathan Leipsic, et al. Diagnosis of ischemia-causing coronary stenoses by noninvasive fractional flow reserve computed from coronary computed tomographic angiograms: results from the prospective multicenter discover-flow (diagnosis of ischemia-causing stenoses obtained via noninvasive fractional flow reserve) study. *Journal of the American College of Cardiology*, 58(19):1989–1997, 2011.

[186] Hovannes Kulhandjian, Prakshi Sharma, Michel Kulhandjian, and Claude D'Amours. Sign language gesture recognition using doppler radar and deep learning. In *2019 IEEE globecom workshops (GC Wkshps)*, pages 1–6. IEEE, 2019.

[187] M. Kunz, D. Meixner, and S. Lautenbacher. Facial muscle movements encoding pain-a systematic review. *Pain*, 160(3):535–549, 2019.

[188] Ilya L. and Frank H. Decoupled weight decay regularization. In *ICLR*, 2019.

[189] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[190] Dmitry Lagun, Cecelia Manzanares, Stuart M Zola, Elizabeth A Buffalo, and Eugene Agichtein. Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of neuroscience methods*, 201(1):196–203, 2011.

[191] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[192] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[193] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[194] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[195] Beichen Li, Jingyu Yang, Yang Yang, Chen Li, and Yutong Zhang. Sign language/gesture recognition based on cumulative distribution density features using uwb radar. *IEEE transactions on instrumentation and measurement*, 70:1–13, 2021.

[196] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.

[197] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[198] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[199] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*, pages 133–141. Springer, 2019.

[200] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[201] Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033, 2020.

[202] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

[203] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection, 2022.

[204] Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025.

[205] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4):1748–1764, 2021.

[206] Feng Lin, Jiarui Han, Teng Xue, Jilan Lin, Shenggen Chen, Chaofeng Zhu, Han Lin, Xianyang Chen, Wanhui Lin, and Huapin Huang. Predicting cognitive impairment in outpatients with epilepsy using machine learning techniques. *Scientific Reports*, 11(1):20002, 2021.

[207] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[208] A.C. Lints-Martindale, T. Hadjistavropoulos, B. Barber, and S.J. Gibson. A psychophysical investigation of the facial action coding system as an index of pain variability among older adults with and without alzheimer's disease. *Pain Med.*, 8(8):678–689, 2007.

[209] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular imaging*, 12(8 Part 1):1549–1565, 2019.

[210] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[211] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[212] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[213] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.

[214] Z. Liu, H. Hu, Y. Lin, Z. Yao, and Z. et al. Xie. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

[215] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[216] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022.

[217] Luca Lo Re. *Corpus multimodale dell'italiano parlato: basi metodologiche per la creazione di un prototipo*. PhD thesis, University of Florence, 2022.

[218] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[219] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[220] Yilong Lu and Yue Lang. Sign language recognition with cw radar and machine learning. In *2020 21st International Radar Symposium (IRS)*, pages 31–34. IEEE, 2020.

[221] Jiahuan Luo, Xueyang Wu, Yun Luo, Anbu Huang, Yunfeng Huang, Yang Liu, and Qiang Yang. Real-world image datasets for federated learning. *arXiv preprint arXiv:1910.11089*, 2019.

[222] Hamzah Luqman, Raffaele Mineo, Murtadha Aljubran, Ahmed Abul Hasanaath, Amelia Sorrenti, Sarah Alyami, Sadam Al-Azani, Maad Alowaifeer, Jihwan Moon, Václav Javorek, et al. The signeval 2025 challenge at the iccv multimodal sign language recognition workshop: Results

and discussion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2025.

[223] David M Lydon-Staley, Ian Barnett, Theodore D Satterthwaite, and Danielle S Bassett. Digital phenotyping for psychiatry: Accommodating data and theory with network science methodologies. *Current opinion in biomedical engineering*, 9:8–13, 2019.

[224] Hua Ma, Pierre Ambrosini, and Theo van Walsum. Fast prospective detection of contrast inflow in x-ray angiograms with convolutional neural network and recurrent neural network. In *MICCAI*, 2017.

[225] Gavin MacLaughlin, Jack Malcolm, and Syed A Hamza. Multi antenna radar system for american sign language (asl) recognition using deep learning. In *Big Data IV: Learning, Analytics, and Applications*, volume 12097, pages 198–204. SPIE, 2022.

[226] Sultan Mohammad Manjur, Sabyasachi Biswas, and Ali C. Gurbuz. A multimodal video and radar fusion framework for high-accuracy isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[227] Mizuki Maruyama, Shrey Singh, Katsufumi Inoue, Partha Pratim Roy, Masakazu Iwamura, and Michifumi Yoshioka. Word-level sign language recognition with multi-stream neural networks focusing on local regions and skeletal information. *IEEE Access*, 2024.

[228] Elliot Mbunge, Benhildah Muchemwa, Sipho'esihle Jiyane, and John Batani. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. *Global Health Journal*, 5(4):169–177, 2021.

[229] James McCleary, Laura Parra García, Christos Ilioudis, and Carmine Clemente. Sign language recognition using micro-doppler and explainable deep learning. In *2021 IEEE Radar Conference (RadarConf21)*, pages 1–6. IEEE, 2021.

[230] R Martin McGuire, Jose Hernandez-Rebollar, Thad Starner, Valerie Henderson, Helene Brashear, and Danielle S Ross. Towards a one-way american sign language translator. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 620–625. IEEE, 2004.

[231] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[232] Z. Meng, S. Han, and Y. Tong. Listen to your face: Inferring facial action units from audio channel. *IEEE Transactions on Affective Computing*, 2017.

[233] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems*, 36, 2024.

[234] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3047–3055, 2017.

[235] Microsoft. Microsoft® Kinect V1 for Windows Sensor.

[236] James K Min, Jonathon Leipsic, Michael J Pencina, Daniel S Berman, Bon-Kwon Koo, Carlos Van Mieghem, Andrejs Erglis, Fay Y Lin, Allison M Dunning, Patricia Apruzzese, et al. Diagnostic accuracy of fractional flow reserve from anatomic ct angiography. *Jama*, 308(12):1237–1245, 2012.

[237] Raffaele Mineo. Advancing healthcare with multimodal ai: Enhanced diagnosis, cognitive monitoring, and inclusive medical communication. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025 Doctoral Symposium*, volume 16489 of *Lecture Notes in Computer Science*. Springer, 2027.

[238] Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Concetto Spampinato, and Simone Palazzo. Learning joint text and visual tokens in clip for medical image analysis. In *Proceedings of the 2nd International Workshop on Multimedia Computing for Health and Medicine*, pages 45–51, 2025.

[239] Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Ulas Bagci, Daniela Giordano, Concetto Spampinato, and Simone Palazzo. Conditional visuo-textual prompt learning for medical image analysis. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.

[240] Raffaele Mineo, Gaia Caligiore, Federica Proietto Salanitri, Isaak Kavasidis, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Radar-based imaging for sign language recognition in medical communication. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2025.

[241] Raffaele Mineo, Gaia Caligiore, Concetto Spampinato, Sabina Fontana, Simone Palazzo, and Egidio Ragonese. Sign language recognition for patient-doctor communication: A multimedia/multimodal dataset. In *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 202–207. IEEE, 2024.

[242] Raffaele Mineo, Isaak Kavasidis, Federica Proietto Salanitri, Lisa Passarello, Giovanni Piccininno, Vincenzo Masciale, Alessandro Anselmo, Cristiano Convertino, Domenico Rotondi, Nicola Laurieri, et al. Automated moca score estimation using eye-gaze data and vision transformers. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 963–968. IEEE, 2025.

[243] Raffaele Mineo, Simone Palazzo, Concetto Spampinato, and Francesco Rundo. Hc-ptq: Poincare-based hyperbolic clustering for data-free quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3994–4001, 2025.

[244] Raffaele Mineo, F Proietto Salanitri, G Bellitto, I Kavasidis, O De Filippo, M Millesimo, GM De Ferrari, M Aldinucci, D Giordano, S Palazzo, et al. A convolutional-transformer model for ffr and ifr assessment from coronary angiography. *IEEE Transactions on Medical Imaging*, 43(8):2866–2877, 2024.

[245] Raffaele Mineo, Federica Proietto Salanitri, Giovanni Bellitto, Ovidio De Filippo, Fabrizio D'Ascenzo, Simone Palazzo, and Concetto Spampinato. Coronary artery stenosis assessment in x-ray angiography through spatio-temporal attention for non-invasive ffr and ifr estimation. In *BIOSTEC (1)*, pages 305–312, 2024.

[246] Raffaele Mineo, Federica Proietto Salanitri, Lisa Passarello, Alberto Sardella, Laura Messina, Manuela Pennisi, Daniela Giordano, Simone Palazzo, and Concetto Spampinato. Magnifying facial micro-movements for cognitive evaluation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024.

[247] Raffaele Mineo, Paolo Soda, Simone Palazzo, and Concetto Spampinato. Biometric safety system and method for estimating a user hazard condition and triggering an emergency stop, 2027. IT Patent.

[248] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. A benchmark for radar-based italian sign language recognition using frequency-domain range-time maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026.

[249] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4894–4902, 2025.

[250] Raffaele Mineo, Amelia Sorrenti, Robin Faro, Gabriele Mineo, Francesco Cancelliere, and Alberto Faro. Pcb-said: A low-cost camera-based dataset for few-shot smd assembly inspection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1351–1357, 2025.

[251] Raffaele Mineo, Amelia Sorrenti, and Federica Proietto Salanitri. Fedetr: A federated approach for stenosis detection in coronary angiography. In *International Conference on Image Analysis and Processing*, pages 189–200. Springer, 2023.

[252] Raffaele Mineo, Paolo Spadaro, Flavio Cannavò, Simone Palazzo, Andrea Cannata, and Concetto Spampinato. Eurovolc-ml/volund-toolkit: Volund v1.0, 2021.

[253] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[254] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pages 19–38. IEEE, 2017.

[255] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450, 2023.

[256] John Mongan, Linda Moy, and Charles E Kahn Jr. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers, 2020.

[257] Marta Montanini, Laura Fruggeri, and Massimo Facchini. *Dal gesto al gesto: il bambino sordo tra gesto e parola*. Cappelli, Bologna, 1979.

[258] Jong Hak Moon, Won Chul Cha, Myung Jin Chung, Kyu-Sung Lee, Baek Hwan Cho, Jin Ho Choi, et al. Automatic stenosis recognition from coronary angiography using convolutional neural networks. *Computer methods and programs in biomedicine*, 198:105819, 2021.

[259] Thamara Carvalho Morais, Antonildes Nascimento Assunção-Jr, Roberto Nery Dantas Júnior, Carla Franco Grego da Silva, Caroline Bastida de Paula, Roberto Almeida Torres, Tiago Augusto Magalhães, César Higa Nomura, Luiz Francisco Rodrigues de Ávila, and José Rodrigues Parga Filho. Diagnostic performance of a machine learning-based ct-derived ffr in detecting flow-limiting stenosis. *Arquivos Brasileiros de Cardiologia*, 116:1091–1098, 2021.

[260] Paul D Morris, Desmond Ryan, Allison C Morton, Richard Lycett, Patricia V Lawford, D Rodney Hose, and Julian P Gunn. Virtual fractional flow reserve from coronary angiography:

modeling the significance of coronary lesions: results from the virtu-1 (virtual fractional flow reserve from coronary angiography) study. *JACC: Cardiovascular Interventions*, 6(2):149–157, 2013.

[261] Manish Motwani, Damini Dey, Daniel S Berman, Guido Germano, Stephan Achenbach, Mouaz H Al-Mallah, Daniele Andreini, Matthew J Budoff, Filippo Cademartiri, Tracy Q Callister, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European heart journal*, 38(7):500–507, 2017.

[262] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

[263] Ayas Faikar Nafis and Nanik Suciati. Sign language recognition on video data based on graph convolutional network. *Journal of Theoretical and Applied Information Technology*, 99(18):4323–4333, 2021.

[264] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks, 2019.

[265] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[266] Franz-Josef Neumann, Miguel Sousa-Uva, AESC Ahlsson, Fernando Alfonso, AP Banning, U Benedetto, et al. Esc scientific document group. 2018 esc/eacts guidelines on myocardial revascularization. *Eur Heart J*, 40(2):87–165, 2019.

[267] Franz-Josef Neumann, Miguel Sousa-Uva, Anders Ahlsson, Fernando Alfonso, Adrian P Banning, Umberto Benedetto, Robert A Byrne, Jean-Philippe Collet, Volkmar Falk, Stuart J Head, Peter Jüni, Adnan Kastrati, Akos Koller, Steen D Kristensen, Josef Niebauer, Dimitrios J Richter, Petar M Seferović, Dirk Sibbing, Giulio G Stefanini, Stephan Windecker, Rashmi Yadav, Michael O Zembala, and ESC Scientific Document Group. 2018 ESC/EACTS Guidelines on myocardial revascularization. *European Heart Journal*, 40(2):87–165, 08 2018.

[268] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[269] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011.

[270] Bjarne L Nørgaard, Jonathon Leipsic, Sara Gaur, Sujith Seneviratne, Brian S Ko, Hiroshi Ito, Jesper M Jensen, Laura Mauri, Bernard De Bruyne, Hiram Bezerra, et al. Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: the nxt trial (analysis of coronary blood flow using ct angiography: Next steps). *Journal of the American College of Cardiology*, 63(12):1145–1155, 2014.

[271] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future-big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[272] Corinna Onelli, Domenico Proietti, Corrado Seidenari, Fabio Tamburini, et al. The diacoris project: a diachronic corpus of written italian. In *LREC*, pages 1212–1215, 2006.

[273] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[274] Antoine Oudin, Redwan Maatoug, Alexis Bourla, Florian Ferreri, Olivier Bonnot, Bruno Millet, Félix Schoeller, Stéphane Mouchabac, and Vladimir Adrien. Digital phenotyping: data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*, 25:e44502, 2023.

[275] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[276] Emmanuel Ovalle-Magallanes, Juan Gabriel Avina-Cervantes, Ivan Cruz-Aceves, and Jose Ruiz-Pinales. Hybrid classical–quantum convolutional neural network for stenosis detection in x-ray coronary angiography. *Expert Systems with Applications*, 189:116112, 2022.

[277] Simone Palazzo, Giovanni Bellitto, Raffaele Mineo, Amelia Sorrenti, Camila Gonzalez, Angelo Porrello, and Lorenzo Bonicelli, editors. *Proceedings of the International Workshop on Personalized Incremental Learning in Medicine (PILM '25), co-located with ACM Multimedia 2025*. Association for Computing Machinery, 2025.

[278] Simone Palazzo, Raffaele Mineo, Concetto Spampinato, Flavio Cannavò, Andrea Cannata, Chiara Montagna, Jean-Marie Saurel, O. Geber, and Arnaud Lemarchand. Machine learning tool kit and database. Deliverable report d9.3, EUROVOLC Project, 2021. Horizon 2020 EUROVOLC, Grant Agreement No. 731070.

[279] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[280] M Panagiotou, A Zlatintsi, PP Filntisis, AJ Roumeliotis, N Efthymiou, and P Maragos. A comparative study of autoencoder architectures for mental health analysis using wearable sensors data. In *EUSIPCO*. IEEE, 2022.

[281] Kun Pang, Danni Ai, Huihui Fang, Jingfan Fan, Hong Song, and Jian Yang. Stenosis-detnet: Sequence consistency-based stenosis detection for x-ray coronary angiography. *Computerized Medical Imaging and Graphics*, 89:101900, 2021.

[282] Rushi V Parikh, Grace Liu, Mary E Plomondon, Thomas SG Sehested, Mark A Hlatky, Stephen W Waldo, and William F Fearon. Utilization and outcomes of measuring fractional flow reserve in patients with stable ischemic heart disease. *Journal of the American College of Cardiology*, 75(4):409–419, 2020.

[283] Jungwoon Park, Junyoung Jang, Geunhaeng Lee, Hyunmin Koh, Changhwan Kim, and Tae Wook Kim. A time domain artificial intelligence radar for hand gesture recognition using 33-ghz direct sampling. In *2019 Symposium on VLSI Circuits*, pages C24–C25. IEEE, 2019.

[284] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

[285] Matteo Pennisi, Federica Proietto Salanitri, Giovanni Bellitto, Simone Palazzo, Ulas Bagci, and Concetto Spampinato. A privacy-preserving walk in the latent space of generative models for medical applications. In *MICCAI*, 2023.

[286] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford university press, 2003.

[287] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 572–578. Springer, 2015.

[288] Nico HJ Pijls, Bernard de Bruyne, Kathinka Peels, Pepijn H van der Voort, Hans JRM Bonnier, Jozef Bartunek, and Jacques J Koolen. Measurement of fractional flow reserve to assess the functional severity of coronary-artery stenoses. *New England Journal of Medicine*, 334(26):1703–1708, 1996.

[289] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[290] C. Pino, S. Palazzo, F. Trenta, F. Cordero, and U. et al. Bagci. Interpretable deep model for predicting gene-addicted non-small-cell lung cancer in ct scans. In *IEEE ISBI*, 2021.

[291] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[292] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

[293] Federica Proietto Salanitri, Giovanni Bellitto, Ismail Irmakci, Simone Palazzo, Ulas Bagci, and Concetto Spampinato. Hierarchical 3d feature learning forpancreas segmentation. In *MLMI (MICCAI workshop)*, 2021.

[294] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.

[295] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[296] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[297] M Mahbubur Rahman, Robiulhossain Mdrafi, Ali C Gurbuz, Evie Malaia, Chris Crawford, Darrin Griffin, and Sevgi Z Gurbuz. Word-level sign language recognition using linguistic adaptation of 77 ghz fmcw radar data. In *2021 IEEE Radar Conference (RadarConf21)*, pages 1–6. IEEE, 2021.

[298] Mohammad Mahbubur Rahman, Evie A Malaia, Ali Cafer Gurbuz, Darrin J Griffin, Chris Crawford, and Sevgi Zubeyde Gurbuz. Effect of kinematics and fluency in adversarial synthetic data generation for asl recognition with rf sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4):2732–2745, 2022.

[299] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[300] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[301] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[302] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[303] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[304] Dharma Rane, Deba Prasad Dash, Alakananda Dutt, Anirban Dutta, Abhijit Das, and Uttama Lahiri. Distinctive visual tasks for characterizing mild cognitive impairment and dementia using oculomotor behavior. *Frontiers in Aging Neuroscience*, 15:1125651, 2023.

[305] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[306] Sunitha Ravi, Maloji Suman, PVV Kishore, Kiran Kumar, Anil Kumar, et al. Multi modal spatio temporal co-trained cnns with single modal testing on rgb–d based sign language gesture recognition. *Journal of Computer Languages*, 52:88–102, 2019.

[307] G. Ravipati, W. S Aronow, H. Lai, J. Shao, A. J DeLuca, M. B Weiss, Anthony L Pucillo, Kumar Kalapatapu, Craig E Monsen, and Robert N Belkin. Comparison of sensitivity, specificity, positive predictive value, and negative predictive value of stress testing versus 64-multislice coronary computed tomography angiography in predicting obstructive coronary artery disease diagnosed by coronary angiography. *The American journal of cardiology*, 101, 2008.

[308] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[309] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–13, 2016.

[310] Aihu Ren, Yong Wang, Xiaobo Yang, and Mu Zhou. A dynamic continuous hand gesture detection and recognition method with fmcw radar. In *2020 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 1208–1213. IEEE, 2020.

[311] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[312] G Retsinas, P P Filntisis, N Efthymiou, E Theodosis, A Zlatintsi, and P Maragos. Person identification using deep convolutional neural networks on short-term signals from wearable sensors. In *ICASSP*. IEEE, 2020.

[313] Flavia Riccardo, Marco Ajelli, Xanthi D Andrianou, Antonino Bella, Martina Del Manso, Massimo Fabiani, Stefania Bellino, Stefano Boros, Alberto Mateo Urdiales, Valentina Marziano, et al. Epidemiological characteristics of covid-19 cases and estimates of the reproductive numbers 1 month into the epidemic, italy, 28 january to 31 march 2020. *Eurosurveillance*, 25(49):2000790, 2020.

[314] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[315] Richard D Riley, Danielle van der Windt, Peter Croft, and Karel GM Moons. *Prognosis research in healthcare: concepts, methods, and impact*. Oxford University Press, 2019.

[316] Dinis L Rodrigues, Miguel Nobre Menezes, Fausto J Pinto, and Arlindo L Oliveira. Automated detection of coronary artery stenosis in x-ray angiography using deep neural networks. *arXiv preprint arXiv:2103.02969*, 2021.

[317] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[318] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[319] F Proietto Salanitri, Giovanni Bellitto, Simone Palazzo, Ismail Irmakci, M Wallace, C Bolan, Megan Engels, Sanne Hoogenboom, Marco Aldinucci, Ulas Bagci, et al. Neural transformers for intraductal papillary mucosal neoplasms (ipmn) classification in mri images. In *EMBC*, 2022.

[320] Federica Proietto Salanitri, Giovanni Bellitto, Raffaele Mineo, Matteo Pennisi, Amelia Sorrenti, Salvatore Calcagno, Daniela Giordano, Simone Palazzo, and Concetto Spampinato. Dynamic graph attention: unraveling spatio-temporal synchrony in eeg data. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2201–2206. IEEE, 2023.

[321] Mirko Santoro and Fabio Poletti. L'annotazione del corpus. In *Grammatica, lessico e dimensioni di variazione nella LIS*, pages 69–78. Franco Angeli, 2011.

[322] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[323] Dmitriy Sazonov, Kamrul Islam, Evie Malaia, and Sevgi Gurbuz. Modality-specific benchmarks and radar range-doppler envelope classification for multimodal isolated sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[324] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[325] Sayan Sen, Javier Escaned, Iqbal S Malik, Ghada W Mikhail, Rodney A Foale, Rafael Mila, Jason Tarkin, Ricardo Petraco, Christopher Broyd, Richard Jabbour, et al. Development and validation of a new adenosine-independent index of stenosis severity from coronary wave–intensity analysis: results of the advise (adenosine vasodilator independent stenosis evaluation) study. *Journal of the American College of Cardiology*, 59(15):1392–1402, 2012.

[326] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[327] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical image analysis*, 88:102802, 2023.

[328] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[329] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

[330] ShiYing Shen, Wenhao Qi, Jianwen Zeng, Sixie Li, Xin Liu, Xiaohong Zhu, Chaoqun Dong, Bin Wang, Yankai Shi, Jiani Yao, et al. Passive sensing for mental health monitoring using machine learning with wearables and smartphones: Scoping review. *Journal of medical Internet research*, 27:e77066, 2025.

[331] Jianbo Shi and Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[332] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[333] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[334] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[335] Mohammad Shorfuzzaman and M Shamim Hossain. Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients. *Pattern recognition*, 113:107700, 2021.

[336] Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

[337] Ala addin I Sidig, Hamzah Luqman, and Sabri A Mahmoud. Arabic sign language recognition using optical flow-based features and hmm. In *Recent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT 2017)*, pages 297–305. Springer, 2018.

[338] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[339] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE access*, 8:181340–181355, 2020.

[340] K Singh, BK Jacobsen, S Solberg, KH Bønaa, S Kumar, R Bajic, and E Arnesen. Intra-and interobserver variability in the measurements of abdominal aortic and common iliac artery diameter with computed tomography. the tromsø study. *European journal of vascular and endovascular surgery*, 25(5):399–407, 2003.

[341] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[342] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 1961.

[343] Alessia Spitalleri, Isaak Kavasidis, Vincenzo Cartelli, Raffaele Mineo, Francesco Rundo, Simone Palazzo, Concetto Spampinato, and Daniela Giordano. Biotrak: A blockchain-based platform for food chain logistics traceability. In *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*, pages 105–110. IEEE, 2023.

[344] Maximilian Springenberg, Annika Frommholz, Markus Wenzel, Eva Weicken, Jackie Ma, and Nils Strodthoff. From cnns to vision transformers – a comprehensive evaluation of deep learning models for histopathology, 2022.

[345] Harshinee Sriram, Cristina Conati, and Thalia Field. Classification of alzheimer's disease with deep learning on eye-tracking data. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 104–113, 2023.

[346] Steven R Steinhubl, Evan D Muse, and Eric J Topol. The emerging field of mobile health. *Science translational medicine*, 7(283):283rv3–283rv3, 2015.

[347] Stereolabs. Stereolabs® ZED 2 AI Stereo Cameras.

[348] Ewout W Steyerberg. Applications of prediction models. In *clinical prediction models: a practical approach to development, validation, and updating*, pages 11–31. Springer, 2008.

[349] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

[350] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, 2015.

[351] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[352] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[353] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

[354] Mingxing Tan and Quoc V Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[355] Matteo Tebaldi, Simone Biscaglia, Massimo Fineschi, Giuseppe Musumeci, Alfredo Marchese, Antonio Maria Leone, Marco Luciano Rossi, Giulio Stefanini, Antongiulio Maione, Alberto Menozzi, et al. Evolving routine standards in invasive hemodynamic assessment of coronary stenosis: the nationwide italian sici-gise cross-sectional eris study. *JACC: Cardiovascular Interventions*, 11(15):1482–1491, 2018.

[356] Bernardo Camajori Tedeschini, Stefano Savazzi, Roman Stoklasa, Luca Barbieri, Ioannis Stathopoulos, Monica Nicoli, and Luigi Serio. Decentralized federated learning for healthcare networks: A case study on tumor segmentation. *IEEE Access*, 10:8693–8708, 2022.

[357] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 151–160. Springer, 2022.

[358] Elena Tomasuolo, Tiziana Gulli, Virginia Volterra, and Sabina Fontana. The italian deaf community at the time of coronavirus. *Frontiers in Sociology*, 5:612559, 2021.

[359] Pim AL Tonino, Bernard De Bruyne, Nico HJ Pijls, Uwe Siebert, Fumiaki Ikeno, Marcel vant Veer, Volker Klauss, Ganesh Manoharan, Thomas Engstrøm, Keith G Oldroyd, et al. Fractional flow reserve versus angiography for guiding percutaneous coronary intervention. *New England Journal of Medicine*, 360(3):213–224, 2009.

[360] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[361] Sandrine Tornay, Oya Aran, and Mathew Magimai-Doss. An hmm approach with inherent model selection for sign language and gesture recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6049–6056, 2020.

[362] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[363] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[364] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[365] Sieu Tran. Generalizable sign language recognition via local temporal convolutions and region-aware pose encoding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[366] Sieu Tran. Region-aware pose modeling and permutation decoding for signer independent sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[367] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[368] Elahe Vahdani, Longlong Jing, Matt Huenerfauth, and Yingli Tian. Multi-modal multi-channel american sign language recognition. *International Journal of Artificial Intelligence and Robotics Research*, 1(01):2450001, 2024.

[369] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

[370] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230, 2019.

[371] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[372] Baptiste Vasey, David A. Clifton, Gary S. Collins, Alastair K. Denniston, Livia Faes, Bart F. Geerts, Xiaoxuan Liu, Lauren Morgan, Peter Watkinson, and Peter McCulloch. Decide-ai: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine*, 27(2):186–187, 2021.

[373] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[374] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[375] Virginia Volterra. *I segni come le parole: la comunicazione dei sordi*. Boringhieri, Torino, 1981.

[376] Virginia Volterra, Maria Roccaforte, Alessio Di Renzo, and Sabina Fontana. *Italian Sign Language from a Cognitive and Socio-semiotic Perspective: Implications for a general language theory*. John Benjamins, 2022.

[377] Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE, 2008.

[378] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[379] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013.

[380] Tao Wan, Hongxiang Feng, Chao Tong, Deyu Li, and Zengchang Qin. Automated identification and grading of coronary artery stenoses with x-ray angiography. *Computer methods and programs in biomedicine*, 167:13–22, 2018.

[381] Pichao Wang, Wanqing Li, Song Liu, Zhimin Gao, Chang Tang, and Philip Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 7–12. IEEE, 2016.

[382] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[383] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018.

[384] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[385] Yong Wang, Aihu Ren, Mu Zhou, Wen Wang, and Xiaobo Yang. A novel detection and recognition method for continuous hand gesture using fmcw radar. *IEEE Access*, 8:167264–167275, 2020.

[386] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[387] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023.

[388] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[389] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[390] Robert F Wolff, Karel GM Moons, Richard D Riley, Penny F Whiting, Marie Westwood, Gary S Collins, Johannes B Reitsma, Jos Kleijnen, Sue Mallett, and PROBAST Group†. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58, 2019.

[391] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[392] H.Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):65:1–65:8, 2012.

[393] Wei Wu, Jingyang Zhang, Hongzhi Xie, Yu Zhao, Shuyang Zhang, and Lixu Gu. Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint. *Computers in biology and medicine*, 118:103657, 2020.

[394] Zhaoyang Xia, Yixiang Luomei, Chenglong Zhou, and Feng Xu. Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):4749–4764, 2020.

[395] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[396] Wufeng Xue, Gary Brahm, Sachin Pandey, Stephanie Leung, and Shuo Li. Full left ventricle quantification via deep multitask relationships learning. *Medical image analysis*, 43:54–65, 2018.

[397] April Yujie Yan, Traci Jenelle Speed, and Casey Overby Taylor. Relapse prediction using wearable data through convolutional autoencoders and clustering for patients with psychotic disorders. *Scientific Reports*, 15(1):18806, 2025.

[398] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu. How fast are the leaked facial expressions: the duration of micro-expressions. *J. Nonverb. Behav.*, 37:217–230, 2013.

[399] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[400] Yifan Yang, Peiqi Jiao, Zixi Nan, and Xilin Chen. A closer look at skeleton-based continuous sign language recognition. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

[401] Liping Yi, Jinsong Zhang, Rui Zhang, Jiaqi Shi, Gang Wang, and Xiaoguang Liu. Su-net: an efficient encoder-decoder model of federated learning for brain tumor segmentation. In *International Conference on Artificial Neural Networks*, pages 761–773. Springer, 2020.

[402] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[403] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[404] Peihua Yu and Yunfeng Liu. Federated object detection: Optimizing object detection model with federated learning. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, pages 1–6, 2019.

[405] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[406] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022.

[407] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

[408] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[409] Dong Zhang, Guang Yang, Shu Zhao, Yanping Zhang, Dhanjoo Ghista, Heye Zhang, and Shuo Li. Direct quantification of coronary artery stenosis through hierarchical attentive multi-view learning. *IEEE Transactions on Medical Imaging*, 39(12):4322–4334, 2020.

[410] Dong Zhang, Guang Yang, Shu Zhao, Yanping Zhang, Heye Zhang, and Shuo Li. Direct quantification for coronary artery stenosis using multiview learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 449–457. Springer, 2019.

[411] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[412] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[413] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[414] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *MICCAI*. Springer, 2021.

[415] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[416] Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. Text-guided foundation model adaptation for pathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 272–282. Springer, 2023.

[417] Chen Zhao, Haipeng Tang, Daniel McGonigle, Zhuo He, Chaoyang Zhang, Yu-Ping Wang, Hong-Wen Deng, Robert Bober, and Weihua Zhou. A new approach to extracting coronary arteries and detecting stenosis in invasive coronary angiograms. *arXiv preprint arXiv:2101.09848*, 2021.

[418] Chen Zhao, Aviral Vij, Saurabh Malhotra, Jinshan Tang, Haipeng Tang, Drew Pienta, Zhihui Xu, and Weihua Zhou. Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms. *Computers in Biology and Medicine*, 136:104667, 2021.

[419] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*, 2021.

[420] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[421] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Cocoop: Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022.

[422] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[423] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[424] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[425] Qidan Zhu, Jing Li, Fei Yuan, and Quan Gan. Continuous sign language recognition based on motor attention mechanism and frame-level self-distillation. *Machine Vision and Applications*, 36(1):1–12, 2025.

[426] Frederik M Zimmermann, Angela Ferrara, Nils P Johnson, Lokien X Van Nunen, Javier Escaned, Per Albertsson, Raimund Erbel, Victor Legrand, Hyeong-Cheol Gwon, Wouter S Remkes, et al. Deferral vs. performance of percutaneous coronary intervention of functionally non-significant coronary stenosis: 15-year follow-up of the defer trial. *European heart journal*, 36(45):3182–3188, 2015.

[427] A Zlatintsi, P P Filntisis, C Garoufis, N Efthymiou, P Maragos, A Menychtas, I Maglogiannis, et al. E-prevention: Advanced support system for monitoring and relapse prevention in patients with psychotic disorders analyzing long-term multimodal data from wearables and video captures. *Sensors*, 22(19):7544, 2022.