ID N. 18799

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVIII Cycle

Enhancing Predictive Models in Lung Cancer with Unstructured EHR and Multi-Modal Data Fusion

Supervisors

Alessandro Bria
Rosa Sicilia
Sara Ramella

Candidate

Domenico Paolo

March, 2026

To my family and friends, who always believed in me.

Acknowledgements

This work is the result of my efforts throughout the PhD period; however, it would not have been possible without the support of many people. I am deeply grateful not only to those who directly contributed to and supported the activities related to this research, but also to those who sincerely supported, and patiently endured, me during these three years. Without each of these contributions, whether small or significant, this work would not have been possible in its present form. While it is impossible to acknowledge everyone individually, I wish to express my deepest gratitude to all those not explicitly mentioned here who, in one way or another, supported me, even through a few words of encouragement or a genuine, kind, and selfless smile.

I would like to express my sincere gratitude to my supervisors, Alessandro Bria and Rosa Sicilia, for their guidance, patience, and invaluable support throughout this PhD journey. They consistently provided insightful suggestions, clear directions, and valuable advice that greatly contributed to my academic growth and to the improvement of this work.

I also thank the members of the ArCo research group at Campus Bio-Medico University for the stimulating discussions and the collaborative environment. In particular, I am grateful to Prof. Paolo Soda for giving me the freedom and the tools necessary to successfully face the daily challenges of my research.

A special thanks also goes to my colleagues at the University of Cassino for welcoming me as part of their group and for their support and patience throughout these three years.

Finally, and most importantly, I thank my family, my brother, and my friends for their constant encouragement, understanding, and unwavering support during these years.

Abstract

Non-small cell lung cancer (NSCLC) remains one of the leading causes of cancer-related mortality worldwide, largely due to its biological heterogeneity, late-stage diagnosis, and limited accuracy of current prognostic models. Reliable prediction of survival and treatment response is essential for personalized therapeutic decision. However, traditional approaches mostly rely on structured clinical variables and fail to capture the rich contextual information embedded in the unstructured components of electronic health records (EHR), such as free-text clinical notes, as well as in medical imaging. In this thesis, we investigate how unstructured clinical text and multimodal data fusion can be leveraged to enhance prognostic modeling in NSCLC and related pulmonary diseases, with particular attention to pulmonary embolism (PE). PE represents a significant contributor to morbidity and mortality among lung cancer patients, and its timely prediction and management are critical for improving clinical outcomes.

First, we develop a domain-specific Named Entity Recognition (NER) system tailored to lung cancer, introducing a clinical ontology and fine-tuning a transformer-based model to extract relevant entities from free-text narratives. Building on these structured semantic representations, we propose a hierarchical attention-based framework that learns patient-level embeddings from clinical notes and demonstrates improved performance in overall survival (OS) and pathologic complete response (pCR) prediction compared to models using only structured data.

To incorporate imaging information, we design a slice-level attention mechanism that processes chest CT scans through 2D convolutional encoders and aggregates prognostic features across slices. This architecture achieves competitive performance on publicly available NSCLC datasets and provides interpretability by highlighting clinically relevant regions.

Finally, we propose a multimodal fusion framework integrating unstructured text, imaging, and structured EHR data. Our fusion strategies (early, late, and attention-based cross-modal integration) consistently outperform unimodal baselines across multiple prediction horizons, when applied to the INSPECT dataset for PE prognosis.

Extensive experiments, ablation studies, and statistical analyses confirm the effectiveness

of our approaches and highlight the complementary nature of clinical text and imaging. This thesis demonstrates that integrating unstructured EHR with multimodal deep learning significantly improves prognostic accuracy and model interpretability, paving the way toward more personalized and data-driven decision support systems in lung cancer care. Future directions include large-scale validation, integration of longitudinal and genomic information, and clinical deployment of multimodal AI systems.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

AI Artificial Intelligence

ALK Anaplastic Lymphoma Kinase

AUC Area Under the Curve

AUPRC Area Under the Precision-Recall Curve

BiLSTM Bidirectional Long-Short Term Memory

BioBIT Biomedical BERT for ITalian

C-index Concordance Index

C^{td}**-index** Time-dependent Concordance Index

CNN Convolutional Neural Network

CNV Copy Number Variation

COPD Chronic Obstructive Pulmonary Disease

CPT Current Procedural Terminology

CSN Cause-Specific sub-Network

CT Computed Tomography

CTk Class Token

CTPA Computed Tomography Pulmonary Angiogram

DFS Disease-Free Survival

DL Deep Learning

EGRF Epidermal Growth Factor Receptor

EHR Electronic Health Record

FDR False Discovery Rate

FC Fully Connected

FN False Negatives

FP False Positives

FPR False Positive Rate

GBM Gradient Boosting Model

HEAL Hierarchical Embedding Attention for overall survivaL

IAA Inter-annotator agreement

ICD International Classification of Diseases

iRECIST Immune Response Evaluation Criteria In Solid Tumors

KRAS Kirsten Rat Sarcoma Viral Oncogene Homolog

LDCT Low-Dose Computed Tomography

MCC Matthews Correlation Coefficient

ML Machine Learning

MLP Multi-Layer Perceptron

NER Named-Entity Recognition

NGS Next-Generation Sequencing

NLP Natural Language Processing

NSCLC Non-Small Cell Lung Cancer

OS Overall Survival

PD-L1 Programmed Death-Ligand 1

PE Pulmonary Embolism

PESI Pulmonary Embolism Severity Index

PFS Progression-Free Survival

pCR Pathologic Complete Response

RNA-seq RNA sequencing (transcriptomic profiling)

SCLC Small Cell Lung Cancer

SN Shared sub-Network

TCIA The Cancer Imaging Archive

TN True Negatives

TNM Tumor, Node, Metastasis (staging system)

TP True Positives

TPR True Positive Rate

TTR Time to Recurrence

TSVD Truncated Singular Value Decomposition

WES Whole-Exome Sequencing

WSI Whole Slide Image

Chapter 1

Introduction

Lung cancer remains one of the leading causes of cancer-related morbidity and mortality worldwide, accounting for approximately 1.8 million deaths annually [145]. Major risk factors include tobacco smoking, environmental exposures (e.g., radon and asbestos), and genetic predispositions, all of which contribute significantly to disease onset and progression [67].

Lung cancer is broadly classified into two main types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), with NSCLC accounting for approximately 85% of all cases. NSCLC includes several subtypes, most notably adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Among these, adenocarcinoma is the most prevalent form, particularly among non-smokers. Despite advances in diagnostic imaging, molecular profiling, and targeted therapies, the overall prognosis for NSCLC patients remains poor. This is largely due to the disease's intrinsic biological and clinical heterogeneity, which complicates accurate prognostication and personalized treatment planning.

Developing reliable prognostic models is therefore crucial for supporting clinical decision-making in NSCLC, especially in guiding individualized therapeutic strategies. In recent years, artificial intelligence (AI) has shown considerable promise in enhancing the accuracy and robustness of prognostic models in NSCLC [99]. However, many traditional AI approaches predominantly rely on structured clinical data, such as demographic information, tumor staging, and molecular markers, which, although informative, may not capture the full complexity and nuance of patients' health as documented in real-world clinical environments. The growing availability of Electronic Health Records (EHRs), especially unstructured data such as physician notes, radiology reports, and discharge summaries, provides an underutilized but promising resource for improving prognostic models. When combined with imaging and other structured data, these unstructured sources offer a comprehensive view of patient health status.

This thesis explores how unstructured EHR data, in conjunction with clinical and imaging

modalities, and multimodal data fusion techniques, can enhance AI predictive models for NSCLC. By integrating free-text clinical narratives with structured clinical variables and imaging-derived information, we aim to improve prognosis in NSCLC patients using deep learning approaches [85].

In this chapter an overview of the broader clinical and technological context in which this research is provided. The chapter begins with a brief overview of NSCLC, including its subtypes, risk factors, and progression patterns, followed by a detailed discussion of the current challenges in prognosis. Subsequently, the motivation for leveraging unstructured and multimodal data in predictive modeling is introduced and the most recent techniques proposed in literature are reviewed. Finally, an outline of the thesis is presented.

1.1 The Non-Small Cell Lung Cancer (NSCLC)

NSCLC is the most common form of lung cancer, accounting for approximately 85% of all lung cancer cases. It also remains the leading cause of cancer-related deaths globally [66]. NSCLC differs from SCLC, the other major subtype of lung cancer, in several key aspects, particularly in histological appearance and biological behavior (Figure 1.1). Under the microscope, NSCLC cells are generally larger and exhibit slower growth and metastasis compared to the more aggressive SCLC.

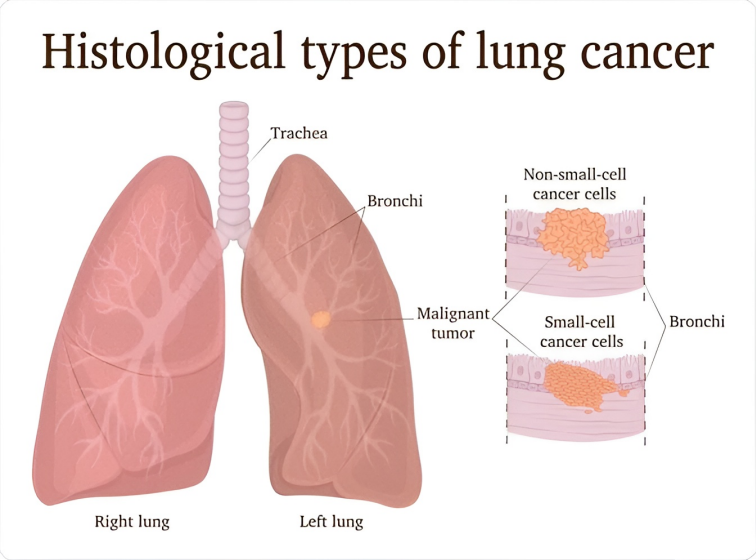

Figure 1.1: Histological differences between NSCLC and SCLC. Source: https://www.screenmaine.org/lung-cancer

As illustrated in Figure 1.2, NSCLC encompasses a diverse group of histological subtypes,

each characterized by distinct morphological and clinical features. The most prevalent subtypes include:

- **Adenocarcinoma**: Arising predominantly in the peripheral regions of the lungs, this is the most common subtype of NSCLC, particularly among non-smokers and younger patients. It is often associated with specific genetic mutations such as EGFR and ALK rearrangements, which have important therapeutic implications.
- **Squamous cell carcinoma**: Typically developing in the central parts of the lungs near the bronchial tubes, this subtype is strongly linked to tobacco exposure. It is characterized by keratinization and the presence of intercellular bridges on histology.
- **Large cell carcinoma**: A less frequent but more aggressive form, large cell carcinoma can arise in any region of the lungs. It is a diagnosis of exclusion, defined by the absence of features typical of adenocarcinoma or squamous cell carcinoma, and is often associated with rapid progression and poor prognosis.

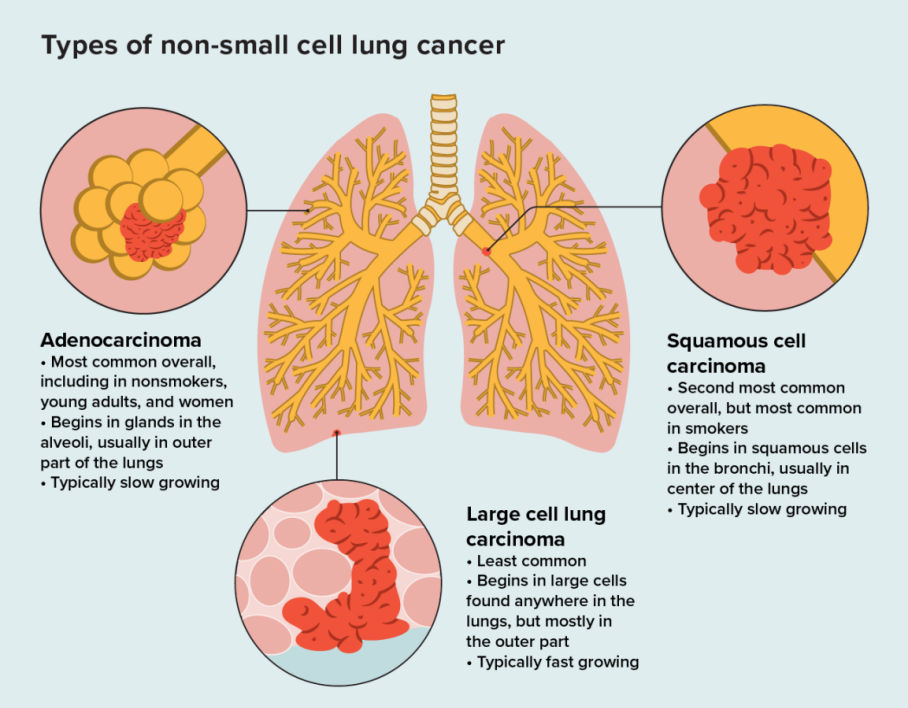

Figure 1.2: Schematic representation of the three major histological subtypes of NSCLC: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Each subtype varies in terms of location, morphology, and associated risk factors. Source: https://www.healthline.com/health/lung-cancer/types-of-non-small-cell-lung-cancer

A wide range of risk factors has been implicated in the development of NSCLC. Tobacco smoking remains the most significant, accounting for the majority of cases; however, other

contributors are increasingly recognized, especially among non-smokers. Environmental exposures such as radon gas, asbestos, air pollution, and occupational carcinogens (e.g., arsenic, chromium, and diesel exhaust) have been strongly associated with increased risk [162]. In addition, genetic susceptibility and a family history of lung cancer may predispose individuals to the disease. Recent studies have also pointed to the role of chronic inflammatory lung conditions, such as chronic obstructive pulmonary disease (COPD), in lung carcinogenesis [157].

The progression of NSCLC is typically characterized by a multistep process involving genetic mutations, abnormal cell proliferation, and eventual invasion of surrounding tissues. Common molecular alterations include EGFR mutations, ALK rearrangements, KRAS mutations, and PD-L1 overexpression, which not only drive tumor growth but also influence treatment response and prognosis. The disease often remains asymptomatic in its early stages, leading to delayed diagnosis and reduced survival rates. When symptoms do appear, such as persistent cough, chest pain, hemoptysis, or unexplained weight loss, they are often indicative of advanced-stage disease.

NSCLC tends to metastasize through lymphatic and hematogenous routes, frequently spreading to the brain, bones, liver, and adrenal glands. The heterogeneous nature of the disease results in highly variable progression patterns among patients, making prognosis challenging. This heterogeneity underscores the urgent need for advanced predictive tools that can capture subtle variations in clinical, radiological, and pathological profiles to inform personalized management strategies.

1.2 Clinical and Prognostic Challenges in NSCLC

As introduced in the previous chapter, NSCLC presents a multifaceted clinical and biological challenge that undermines accurate prognostication. To fully appreciate the challenges and opportunities in developing robust prognostic models for NSCLC, it is important to first examine the clinical complexity of the disease and the limitations of current predictive approaches. Accordingly, this chapter examines the specific clinical burdens, sources of heterogeneity, and current limitations in prognosis that motivate the development of more advanced predictive models.

1.2.1 Clinical Burden of NSCLC

NSCLC imposes a substantial clinical and economic burden on healthcare systems worldwide. As illustrated in Figure 1.3, adapted from [41], the majority of NSCLC cases are

diagnosed at advanced stages of the disease, leading to poor outcomes and limited curative options. Despite progress in screening, diagnosis, and treatment, such as low-dose computed tomography (LDCT) screening programs and the advent of targeted therapies, the five-year overall survival (OS, defined as the proportion of patients alive five years after diagnosis) rate for NSCLC remains below 25% in many regions [135].

This persistently high mortality is driven not only by late-stage diagnosis but also by a range of additional factors, including tumor resistance mechanisms, and significant patient-to-patient variability in response to treatment. Moreover, disparities in access to precision medicine approaches, such as genomic profiling and immunotherapy, further exacerbate the clinical challenge, particularly in low-resource settings.

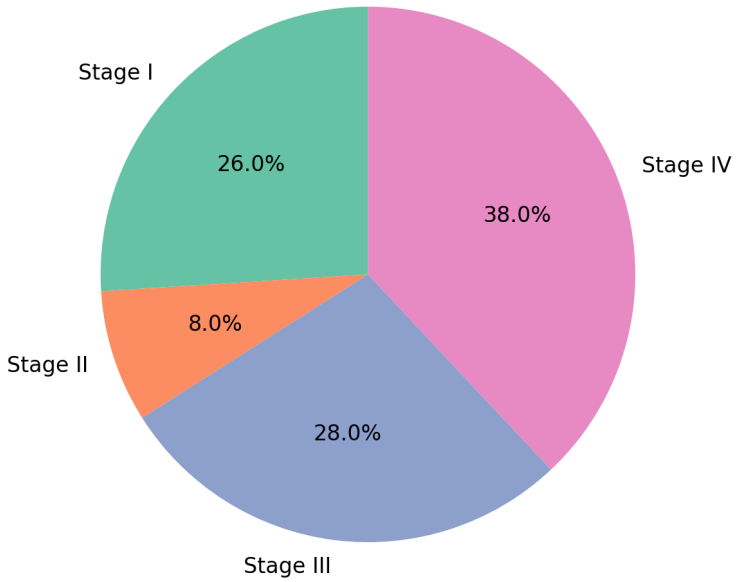

Figure 1.3: Stage distribution of NSCLC at diagnosis. Most patients are diagnosed at stage III or IV, where curative treatment is less likely.

1.2.2 Heterogeneity in NSCLC: Clinical and Biological Dimensions

NSCLC is not a single disease but a biologically and clinically heterogeneous group of malignancies. This heterogeneity manifests at multiple levels:

- **Histological heterogeneity**, with subtypes like adenocarcinoma and squamous cell carcinoma responding differently to therapies.

- **Molecular heterogeneity**, including mutations in EGFR, ALK, KRAS, and other oncogenic drivers, which influence prognosis and therapeutic decisions.
- **Immunological heterogeneity**, reflected in varying PD-L1 expression and tumor-infiltrating immune profiles that affect response to immunotherapy.
- **Radiological and pathological variability**, which impacts staging accuracy and treatment selection.

Such complexity challenges the development of generalizable prognostic models and emphasizes the need for patient-specific approaches. Clinical decisions often rely on a combination of imaging, histology, molecular diagnostics, and clinical judgment, making standardized risk stratification difficult.

1.2.3 Unmet Needs in Prognosis and Treatment

Current clinical guidelines stratify patients into broad risk categories based on stage, histology, and molecular markers. However, these guidelines often fail to capture the dynamic and multifactorial nature of disease progression. Key unmet needs include:

- **More accurate prognostic stratification**, especially within the same stage or treatment group.
- **Better prediction of treatment response**, including neoadjuvant or adjuvant therapies.
- **Identification of early surrogate markers of long-term outcomes**.
- **Integration of longitudinal clinical data**, such as treatment history, laboratory trends, and radiologic evolution, into predictive frameworks.

Traditional statistical models, such as Cox proportional hazards models, are limited in their ability to incorporate high-dimensional, longitudinal, and multimodal data, which are increasingly available through EHRs and real-world datasets.

1.2.4 Clinical Endpoints in Prognostic Modeling

To address the prognostic complexity of NSCLC, various clinical endpoints have been adopted as modeling targets. Among the most widely used are:

Overall Survival (OS)

OS refers to the time from diagnosis (or initiation of treatment) to death from any cause. It remains the gold standard for evaluating treatment efficacy and prognostic models. OS is a definitive and objective outcome but is influenced by a wide range of clinical, therapeutic, and sociodemographic variables. Accurate stage-specific modeling of OS can aid in treatment planning and follow-up intensity.

In the context of AI-driven models, OS presents both an opportunity and a challenge: its clinical relevance is high, but its prediction requires integrating heterogeneous and temporally-evolving patient data.

To evaluate the performance of survival models, the concordance index (C-index) is commonly used. The C-index measures the model's ability to correctly rank pairs of patients based on their predicted survival risk. Specifically, it reflects the probability that, for a randomly selected pair of patients, the patient who dies first had a higher predicted risk score. A C-index of 0.5 indicates random performance, whilst a value of 1.0 reflects perfect concordance between predictions and outcomes. It is particularly useful for assessing models trained on censored survival data, where traditional accuracy metrics are not applicable.

The C-index is formally defined as:

$$C - index = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \tag{1.1}$$

Where:

T_i is the time to event for individual i.

η_i is the predicted risk score for individual i.

δ_j is an indicator function, where $\delta_j = 1$ if the event occurred for individual j and 0 otherwise.

$\mathbf{1}_{T_j<T_i}$ is an indicator function that is 1 if $T_j < T_i$, and 0 otherwise.

$\mathbf{1}_{\eta_j>\eta_i}$ is an indicator function that is 1 if $\eta_j > \eta_i$, and 0 otherwise.

Pathologic Complete Response (pCR)

Pathologic complete response (pCR) is defined as the absence of residual viable tumor cells in resected specimens following neoadjuvant therapy, which is typically administered to patients with resectable, locally advanced NSCLC prior to surgery. Pathologic complete response (pCR) is considered a strong surrogate marker for long-term outcomes such as disease-free survival (DFS), defined as the length of time after treatment during which a patient shows no signs of disease, and OS, defined as the duration of time from diagnosis or start of treatment

until death from any cause, particularly in patients with early-stage or locally advanced NSCLC [42].

Predicting pCR can support therapeutic decision-making by identifying patients likely to benefit from preoperative chemotherapy or immunotherapy. However, due to the complexity of treatment response dynamics, accurate prediction remains a significant challenge.

The performance of predictive models for this binary outcome (pCR vs. non-pCR) is commonly assessed using the Area Under the Receiver Operating Characteristic curve (AUC). This metric measures a model's ability to correctly distinguish between patients who will achieve pCR .

The AUC is formally defined as:

$$\text{AUC} = \frac{1}{|S_P| \cdot |S_N|} \sum_{i \in S_P} \sum_{j \in S_N} \mathbf{1}_{f(x_i) > f(x_j)}$$

Where:

S_P	is the set of positive instances (e.g., patients with pCR).
S_N	is the set of negative instances (e.g., patients without pCR).
$\|S_P\|$	is the number of positive instances.
$\|S_N\|$	is the number of negative instances.
$f(x_i)$	is the predicted score from the model for a positive instance i.
$f(x_j)$	is the predicted score from the model for a negative instance j.
$\mathbf{1}_{f(x_i) > f(x_j)}$	is an indicator function that equals 1 if $f(x_i) > f(x_j)$, and 0 otherwise.

Other Relevant Outcomes

Depending on clinical context, other outcome variables may also be considered, including:

- **Progression-Free Survival (PFS)**, the length of time during and after treatment that a patient lives with the disease without it worsening; often used as an early measure of treatment efficacy.
- **Time to Recurrence (TTR)**, the time from the end of initial treatment to the return of disease, typically used to evaluate long-term control of cancer.
- **Treatment-related toxicity or adverse events**, any negative effects experienced by the patient due to treatment, including side effects from chemotherapy, radiation, or targeted therapies, which may affect quality of life or require treatment modification.

- **Response to immunotherapy (e.g., iRECIST criteria)**, the degree to which a tumor responds to immune-based therapies, measured using standardized guidelines like iRECIST (immune Response Evaluation Criteria In Solid Tumors), which account for unique patterns of response and progression under immunotherapy.

These outcomes, although clinically valuable, are often less standardized in routine EHR documentation, which complicates their use in large-scale predictive modeling.

AI and machine learning (ML) techniques, particularly deep learning, offer powerful tools to navigate the prognostic complexity of NSCLC and to overcome many limitations inherent in traditional statistical models. Despite their potential, however, the clinical adoption of AI-based prognostic models remains limited. Key barriers include insufficient interpretability, poor generalizability across diverse patient populations, and a lack of seamless integration into existing clinical workflows.

This thesis aims to tackle these challenges by developing predictive models that leverage unstructured electronic health record (EHR) narratives and integrate diverse data modalities through multimodal fusion techniques. The objective is to enhance the prediction of key clinical outcomes, such as OS and pCR, in patients with NSCLC.

The following section will describe the various types of data used in NSCLC prognostication, explore the unique challenges associated with each modality, and examine how these data sources can be effectively leveraged to support AI-driven prognostic modeling.

1.3 Multimodal Clinical Data: Opportunities and Challenges

The management and prognostication of NSCLC rely increasingly on a diverse range of clinical data. These data span from structured information stored in EHRs to unstructured clinical narratives in EHRs, radiological imaging, histopathological slides, and genomic profiles. Each modality captures a different aspect of the patient's clinical status, demographics and comorbidities, disease trajectory, tumor burden, and cellular-level pathology, and their integration promises a more accurate and personalized approach to prediction and treatment planning.

Traditional statistical models, although valuable, are often limited by their reliance on a small set of predefined variables and assumptions of linearity or independence. These constraints restrict their ability to capture the complex, nonlinear relationships inherent in high-dimensional clinical data, and they often treat each data modality in isolation.

In contrast, as illustrated in Figure 1.4, recent advances in AI and deep learning have

enabled the development of both unimodal and multimodal models that can learn directly from raw or minimally processed data. Unimodal models focus on extracting deep, task-specific representations from a single data source, such as imaging or text, while multimodal models aim to integrate heterogeneous data types to capture inter-modal interactions and improve predictive accuracy.

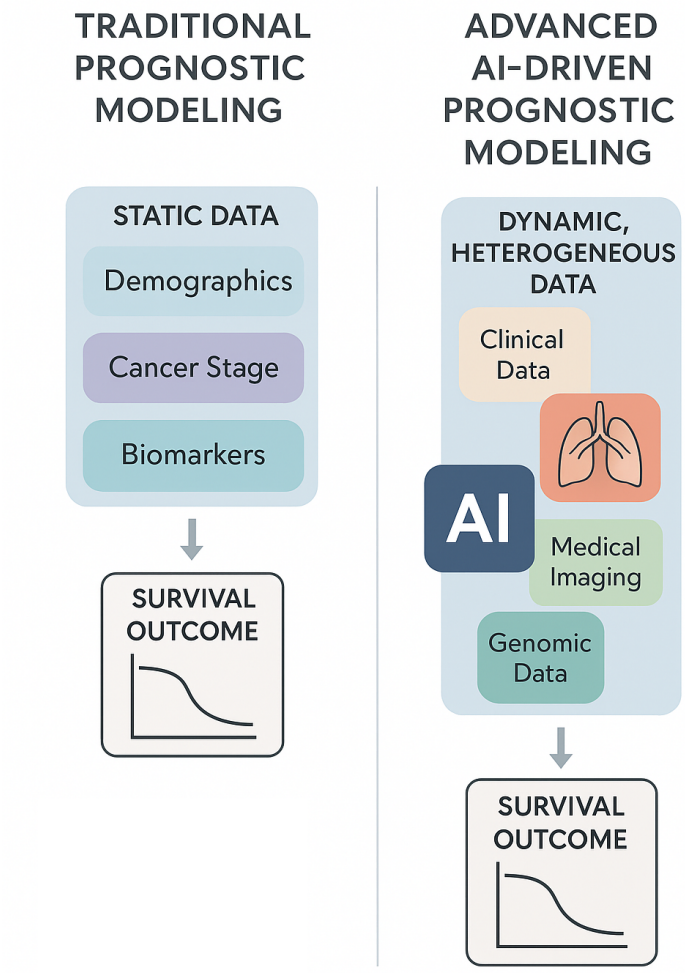

Figure 1.4: Schematic comparison of traditional versus advanced AI-driven prognostic modeling in NSCLC. Traditional models use a small set of static, structured data, whereas AI-based models can learn from dynamic, heterogeneous inputs.

Despite their promise, these AI-driven approaches face significant challenges related to data availability, standardization, alignment across modalities, and clinical interpretability.

This section highlights the unique contributions and limitations of each data modality in

the context of NSCLC, establishing the foundation for the exploration of advanced AI-based prognostic models, both unimodal and multimodal.

1.3.1 Structured EHR

As illustrated in Figure 1.5, structured EHR data consist of encoded patient information such as demographics (e.g., age, sex, ethnicity, smoking status), laboratory test results (e.g., blood counts, liver and kidney function markers), vital signs (e.g., heart rate, oxygen saturation), diagnostic and procedural codes (e.g., ICD, CPT), and medication prescriptions. These elements are typically collected routinely in clinical settings and stored in relational formats that are both accessible and standardized across institutions [160].

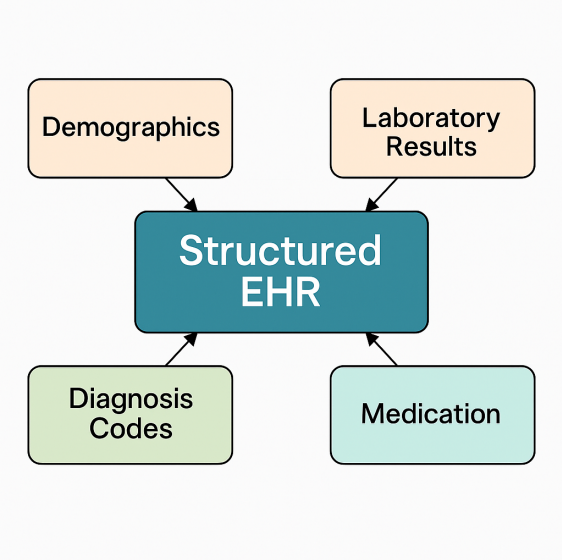

Figure 1.5: Schematic representation of structured EHR data sources. Encoded elements such as demographics, laboratory test results, vital signs, diagnosis codes (ICD), and medication records provide standardized, analyzable patient information crucial for predictive modeling in NSCLC prognosis.

Due to their structured nature, these data are highly amenable to conventional statistical analysis and modern machine learning pipelines. They offer a reliable foundation for training prognostic models and are often used in large-scale epidemiological and clinical research studies. Several transformer-based models have been developed to learn representations from structured EHR data. Med-BERT [129] and BEHRT [93] treat sequences of medical codes as language tokens, modeling patient histories using contextual embeddings. CEHR-BERT [117] and G-BERT [134] extend this with time-aware and graph-enhanced mechanisms, respectively. RETAIN [27] offers interpretable predictions through reverse-time attention.

More recently, MOTOR [141] introduces a modality-aware transformer for multi-task learning across heterogeneous structured EHR data types, improving generalization and efficiency.

However, structured data alone may fail to capture the longitudinal and contextual subtleties essential for nuanced decision-making in NSCLC. For instance, although ICD codes can confirm a diagnosis of lung cancer, they do not convey disease stage, tumor burden, treatment intent (curative vs. palliative), or patient-specific factors such as frailty, functional status, or psychosocial considerations. Moreover, structured fields may miss the clinician's reasoning, temporal patterns in patient trajectories, and subtle changes in clinical status, all of which are vital for precise prognosis and therapeutic planning.

Therefore, although structured EHR data remain an indispensable component of clinical modeling, their limitations underscore the need to integrate richer and more temporally sensitive data modalities, such as free-text narratives and medical imaging, in order to comprehensively model NSCLC progression and outcomes.

1.3.2 Unstructured EHR

Unlike structured EHR fields, unstructured EHR or free-text clinical notes, such as physician observations, discharge summaries, pathology reports, and radiology impressions (Figure 1.6), capture the nuanced and dynamic aspects of patient care that often elude encoded data systems (Figure 1.7).

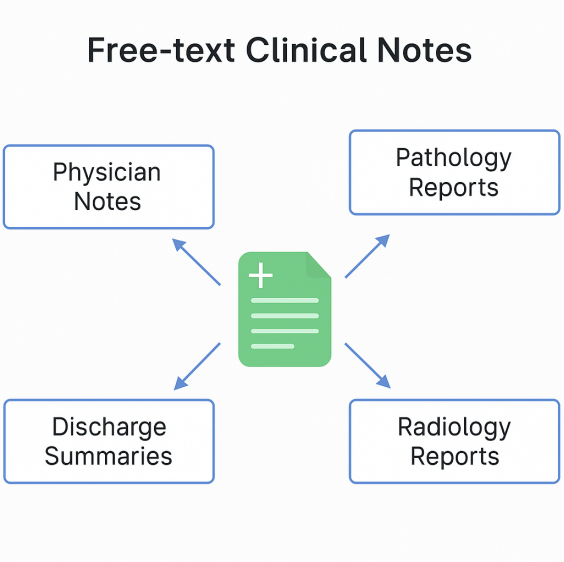

Figure 1.6: Conceptual diagram illustrating the role of free-text clinical notes in EHR. These unstructured narratives, comprising clinician observations, patient histories, and treatment responses, offer rich contextual information that complements structured data, enabling deeper insights into patient trajectories and outcomes.

These narratives reflect not only diagnostic and therapeutic decisions, but also the reasoning processes, uncertainties, and evolving clinical states observed by healthcare providers [44]. Embedded within these documents are critical contextual cues: symptom progression, functional decline, clinician judgment, comorbid conditions, complications, psychosocial factors, and patient preferences. For instance, a pathology report may describe molecular subtypes or ambiguous findings not encoded in the structured record; a radiology impression may note subtle progression suggestive of early recurrence; a physician note might document a patient's refusal of chemotherapy due to side effects, information highly relevant to prognosis, yet difficult to capture in structured fields.

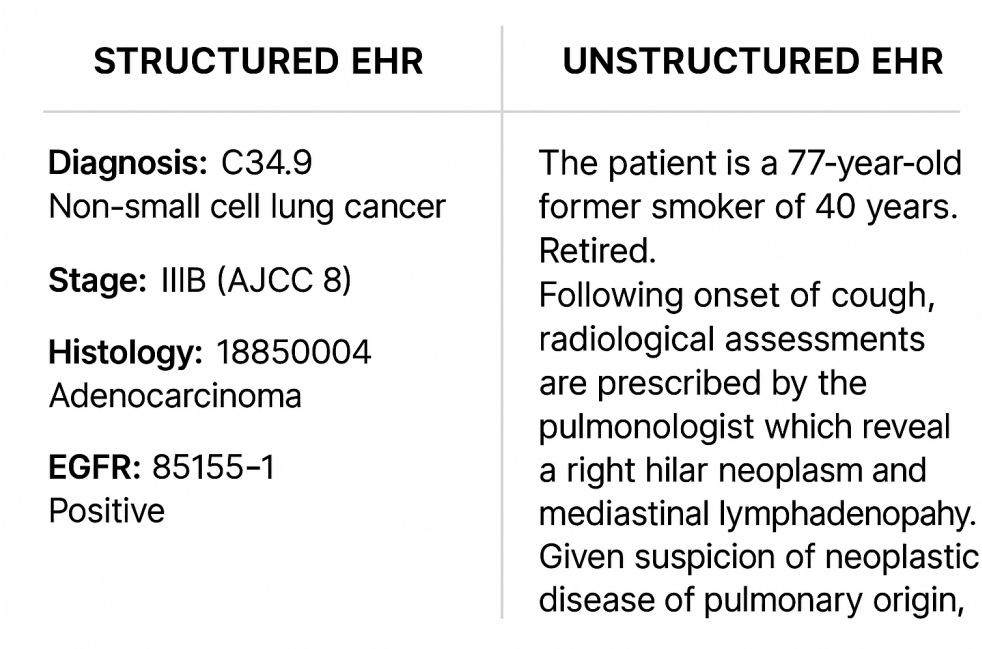

Figure 1.7: Comparison between structured and unstructured EHR in the context of non-small cell lung cancer. The structured format presents codified clinical concepts using standardized terminologies (e.g., ICD codes, staging, histology), while the unstructured format expresses the same information as free-text narrative, providing additional context such as smoking history, clinical suspicion, and physician reasoning.

Natural Language Processing (NLP) has emerged as a transformative tool for extracting and modeling the latent knowledge within unstructured EHRs. Traditional rule-based systems and clinical ontologies are increasingly complemented or replaced by neural language models capable of learning contextual representations of clinical language. Recent advances, including transformer-based models such as ClinicalBERT [62] and BioGPT [100], have shown significant promise in predictive tasks such as mortality, progression-free survival (PFS), and treatment response, outperforming models based solely on structured data [80].

Nevertheless, unstructured data pose substantial methodological and operational challenges ([149]). Clinical language is often fragmented, ambiguous, and institution-specific, with inconsistent terminology and abbreviations. Temporal cues are implicit and sometimes contradictory, requiring sophisticated approaches to resolve sequence and temporality. Moreover, annotated datasets for supervised learning remain limited due to privacy concerns and the cost of expert labeling, making self-supervised and weakly supervised learning strategies increasingly attractive.

Despite these hurdles, free-text clinical notes are indispensable for holistic modeling of patient trajectories in NSCLC. When appropriately processed, they offer deep insights into real-world care patterns and can significantly enhance the performance and interpretability of multimodal AI systems in oncology ([89]).

1.3.3 Computed Tomography (CT)

Computed Tomography (CT) plays a pivotal role in the diagnosis, staging, and longitudinal monitoring of NSCLC. It remains the most widely used imaging modality in routine oncology workflows due to its high spatial resolution, accessibility, and ability to provide comprehensive anatomical detail. As illustrated in Figure 1.8, CT scans offer multiplanar visualizations, axial, coronal, and sagittal, that delineate tumor size, precise localization, lymph node involvement, and the presence of distant metastases, which are central components of TNM (Tumor-Node-Metastasis) staging.

Beyond visual inspection, CT imaging can be quantitatively mined through radiomics, a process that extracts high-dimensional features, such as texture, shape, intensity, and wavelet-based patterns, from segmented tumor regions. These radiomic features have demonstrated prognostic value in several studies, especially when combined with clinical data such as patient demographics, staging information, and treatment histories [95]. Radiomic signatures have been associated with outcomes such as OS [155], PFS [21], and response to chemotherapy or immunotherapy, indicating their potential to support personalized medicine.

However, recent advancements in deep learning have shifted the paradigm from handcrafted radiomic descriptors to end-to-end learning from raw pixel data. Convolutional Neural Networks (CNNs) and other advanced architectures can automatically discover hierarchical imaging biomarkers, often outperforming traditional methods in classification and survival prediction tasks [58]. These models can integrate spatial and textural tumor patterns, peri-tumoral features, and even subtle imaging cues imperceptible to the human eye.

Despite these advances, several technical and translational challenges persist. Image acquisition protocols vary widely across institutions and scanner vendors, leading to het-

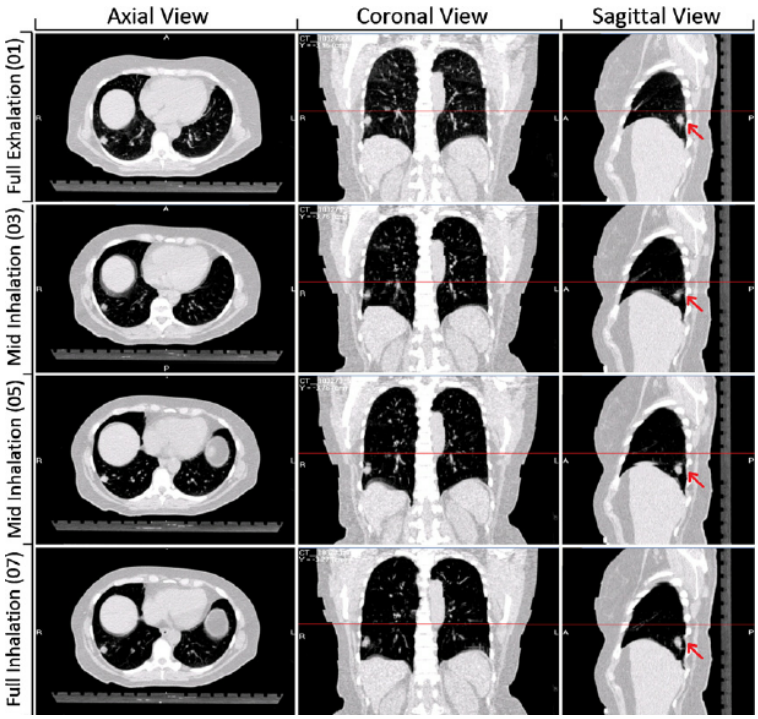

Figure 1.8: Axial, coronal, and sagittal views of a lung tumor as visualized on a CT scan. These multiplanar reconstructions illustrate the tumor's position within the lung, highlighting the importance of 3D imaging for accurate assessment and treatment planning. Source: https://doi.org/10.1088/0031-9155/57/22/7579

erogeneity in resolution, contrast enhancement, and reconstruction algorithms. This lack of standardization can compromise model generalizability. Furthermore, the limited availability of large, annotated imaging datasets, especially with long-term survival labels or treatment outcomes, hinders the training of robust models. Privacy regulations and the time-consuming nature of expert annotation further exacerbate this issue.

Integrating CT-derived features into multimodal prognostic models requires careful preprocessing, harmonization, and validation strategies. When combined with other data modalities, such as EHR-derived variables and histopathology, CT imaging serves as a critical macroscopic layer in the holistic understanding of tumor biology and patient prognosis in NSCLC.

1.3.4 Whole Slide Imaging (WSI)

Histopathology remains the definitive standard for the diagnosis and subtyping of NSCLC. Through tissue biopsy and microscopic examination, pathologists assess key morphological features that define malignancy, grade, and potential therapeutic targets. In the digital era, these glass slides are increasingly converted into whole slide images (WSIs), gigapixel-

resolution scans that preserve cellular-level detail and enable computational analysis [88]. As shown in Figure 1.9, WSIs capture intricate histological structures, including tumor cell density, mitotic activity, nuclear pleomorphism, glandular differentiation, stromal composition, and immune infiltration.

These features are not only diagnostic but also prognostic, as they may correlate with immunotherapy response ([159]). Furthermore, certain histopathological patterns are associated with molecular alterations, such as EGFR mutations or ALK rearrangements, providing indirect cues for molecular testing ([115]).

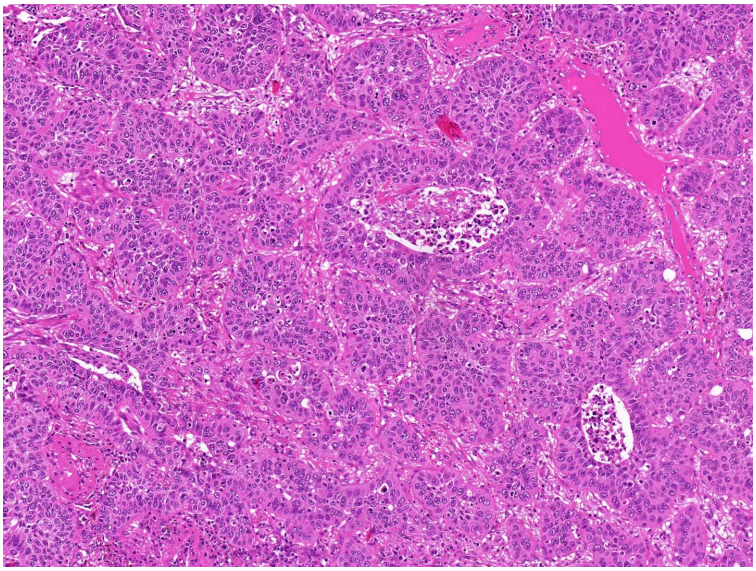

Figure 1.9: Representative WSI of a NSCLC tissue sample. The high-resolution scan enables detailed visualization of tumor architecture, cellular morphology, and histopathological features critical for diagnosis and prognostic assessment. Source: https://www.eurekalert.org/multimedia/659513

Recent advances in deep learning have revolutionized WSI analysis, enabling automated tissue segmentation, tumor classification, grading, and survival prediction. CNNs and transformer-based models can learn discriminative patterns directly from raw pixel data, bypassing the need for hand-crafted features [92, 163]. These models have achieved pathologist-level performance in several benchmark tasks and have shown promise in predicting downstream outcomes, including recurrence and treatment response ([24, 49]).

However, WSIs present unique computational and operational challenges. Each slide can exceed several gigabytes in size, requiring significant storage, memory, and processing capabilities. Annotation at the cellular or regional level is time-consuming and typically relies on expert pathologists, limiting the availability of labeled datasets. Additionally, many pathology departments still rely on analog workflows, and the digitization of slides is not yet

universal across clinical settings. Variability in staining protocols, scanner resolution, and tissue preparation further complicates model generalizability.

Despite these barriers, WSIs offer a rich, underutilized modality for prognostic modeling. When integrated with clinical, imaging, and molecular data, histopathological features can provide ground-truth insights into tumor biology and support more comprehensive, multimodal decision-making frameworks in NSCLC.

1.3.5 Genomic Data: Molecular Insights into NSCLC

Genomic data offer a molecular-level understanding of non-small cell lung cancer (NSCLC), revealing critical alterations that influence disease behavior, therapeutic response, and patient prognosis. These data typically include somatic mutations, copy number variations (CNVs), gene expression profiles, and epigenetic markers derived from tumor tissue or liquid biopsies. Key driver mutations, such as alterations in *EGFR*, *KRAS*, *ALK*, and *TP53*, have been identified as pivotal in the classification and targeted treatment of NSCLC subtypes [161].

As shown in Figure 1.10, genomic information is generally acquired through high-throughput sequencing techniques, including whole-exome sequencing (WES), targeted gene panels, or transcriptomic profiling via RNA sequencing (RNA-seq). These platforms yield high-dimensional data capable of capturing the complex molecular heterogeneity of lung cancer, offering actionable insights for precision medicine strategies.

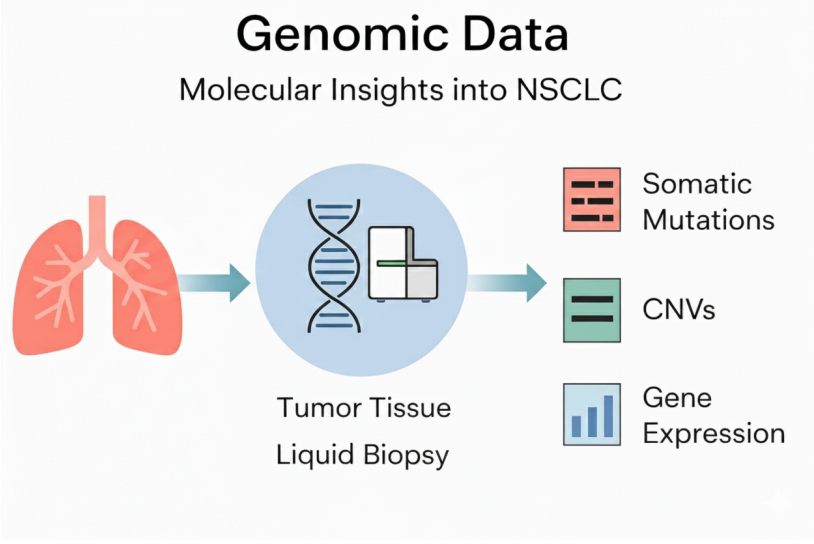

Figure 1.10: Overview of genomic data sources and their role in NSCLC prognosis. Molecular alterations such as somatic mutations, CNVs, and gene expression patterns can inform targeted therapies and survival risk stratification.

In the context of artificial intelligence, genomic data have been successfully used to train predictive models for prognosis and treatment response. For instance, machine learning algorithms have utilized gene expression signatures to classify tumor subtypes and predict clinical outcomes in NSCLC ([120]). Integrative approaches that combine genomic profiles with clinical or imaging data have also demonstrated superior predictive performance compared to unimodal models [32].

Despite their clinical value, genomic data present several challenges. Sequencing costs, variability in assay platforms, and limited availability of annotated genomic datasets across diverse populations restrict their routine implementation in real-world clinical settings. Moreover, the high dimensionality and sparsity of genomic features often necessitate sophisticated dimensionality reduction and regularization strategies to avoid model overfitting ([2]).

Nevertheless, as sequencing becomes more cost-effective and clinically adopted, the integration of genomic data into multimodal AI pipelines is expected to further personalize NSCLC management and improve prognostic accuracy.

1.4 From Unimodal to Multimodal AI

This section reviews advancements in AI applied to clinical data, spanning structured and unstructured EHR data, and examines the trend toward integrating these sources with imaging and molecular data in multimodal models for improved NSCLC prognosis.

1.4.1 AI with Clinical Data: Structured and Unstructured EHR

Structured clinical variables derived from EHRs, including demographic characteristics, TNM staging, vital signs, and laboratory biomarkers, constitute a readily accessible and interpretable data source for AI-based prognostic modeling in NSCLC. Early models typically relied on conventional statistical approaches, such as Cox proportional hazards regression [106] and logistic regression [84], to estimate survival probabilities and clinical outcomes.

With the advent of more sophisticated computational tools, ML techniques have gained traction in this domain. Models such as random forests [130], support vector machines [146], and deep neural networks [85] have been increasingly employed to capture complex, nonlinear relationships among clinical features. For instance, Yuan et al. [164] developed ML-based survival prediction models using structured EHR data from a cohort of over 42,000 lung cancer patients, achieving strong discriminative performance in estimating OS. Similarly, DeepSurv [72], a deep learning–enhanced extension of the Cox model, has demonstrated superior performance compared to traditional survival models in various cancer datasets.

However, despite their utility, models relying exclusively on structured clinical data are inherently limited in their ability to capture nuanced temporal patterns, contextual dependencies, and narrative details embedded in a patient's longitudinal clinical history ([148]). To address this limitation, recent efforts have focused on leveraging unstructured data sources such as free-text clinical notes, discharge summaries, and radiology reports. These data types offer a rich and underutilized reservoir of prognostic information that is often unavailable in encoded formats.

NLP techniques, particularly those based on transformer architectures like BERT [37], have shown promise in extracting clinically relevant signals from these unstructured narratives. For example, Zhou et al. [173] introduced a model incorporating semantic units extracted from unstructured notes using a BERT-based model, which, when combined with radiomic and clinical features, achieved an area under the curve (AUC) of approximately 0.766 in unimodal EHR survival prediction, prior to multimodal fusion.

However, these approaches often lack domain-specific adaptation, and many fail to systematically bridge unstructured and structured modalities ([15, 126]), leaving significant room for improvement in model expressiveness, interpretability, and prognostic performance.

1.4.2 Unimodal Models Using Imaging, Pathology, and Genomics

In addition to structured and unstructured clinical records, unimodal prognostic models have been developed using high-dimensional biomedical data such as radiological imaging, histopathology, and genomic profiles. These modality-specific models leverage advanced feature extraction techniques, both handcrafted and data-driven, to identify imaging- or molecular-based biomarkers that correlate with patient outcomes in NSCLC.

Imaging-based AI models primarily utilize radiomics and deep learning techniques applied to CT or whole-slide histopathology images. Radiomics approaches extract quantitative descriptors related to tumor shape, intensity, and texture, which have been consistently associated with survival across independent NSCLC cohorts [99]. CNN-based pipelines, including models such as LungNet [40], have demonstrated strong AUC performance in survival classification tasks [33]. However, radiomics features often suffer from reproducibility issues due to variations in image acquisition and high dimensionality, leading to potential overfitting.

A key distinction in unimodal imaging-based models pertains to the type of survival outcome being predicted. Many deep learning approaches, particularly those employing standard CNN architectures, frame survival prediction as a binary classification task (e.g., alive/dead at a fixed time point such as 1, 3, or 5 years post-diagnosis) due to the relative

simplicity of optimizing cross-entropy loss and interpreting AUC metrics. While effective for risk stratification at specific thresholds, this binary formulation discards valuable temporal information present in the exact time-to-event data and cannot directly model hazard functions or time-dependent risk.

In contrast, time-to-event (survival) models that explicitly incorporate the temporal dimension, such as Cox proportional hazards extensions (e.g., DeepSurv ([72])), neural network-based parametric survival models, or ranking-based approaches, leverage both censoring status and observed survival times to estimate continuous risk over time. These models are typically evaluated using concordance indices (C-index) rather than AUC alone and offer greater clinical relevance for individualized prognosis.

In the domain of molecular data, genomics- and multi-omics–driven models aim to capture the prognostic relevance of gene expression, DNA methylation, microRNA, and other omic signatures. These models often integrate clinical and molecular data using autoencoder-based architectures or hybrid deep learning frameworks. Reported concordance indices (C-indices) for such models typically range from 0.67 to 0.69 across different NSCLC subtypes [33], underscoring the potential of molecular integration for systems-level prognostication while highlighting limitations in capturing tumor heterogeneity.

Despite these successes, each modality captures only a subset of the multifaceted biological and clinical variability in NSCLC. Imaging alone may reflect tumor morphology but not molecular alterations or treatment response; genomics captures molecular features but not tumor microenvironment or radiographic heterogeneity; clinical notes provide context not available in structured variables. Unimodal approaches are further limited by issues such as overfitting, poor generalizability, and incomplete representation of tumor biology [33]. Consequently, multimodal learning strategies, which integrate clinical, imaging, and molecular data, aim to leverage complementary information from each source. For instance, fusion models combining radiomics, gene expression, and clinical data have demonstrated superior risk stratification, effectively distinguishing low- and high-risk groups with notable accuracy and surpassing traditional unimodal data-based approaches [33]. These integrations enhance prognostic robustness, generalizability, interpretability, and the ability to model time-to-event outcomes [35, 158, 61].

1.4.3 Multimodal Fusion Models

Given the inherent limitations of single-modality models, there is a growing emphasis on multimodal AI frameworks that integrate diverse data sources, such as clinical variables, radiological imaging, histopathology, and molecular profiles, to improve prognostic accuracy

in NSCLC [99]. These models are designed to harness the complementary strengths of different modalities, capturing both macro-level clinical context and micro-level biological signals in a unified analytical framework.

Several recent studies have demonstrated the effectiveness of such multimodal approaches. Zhou et al. [173] proposed a hybrid model that integrates CT radiomics, structured clinical variables, and unstructured EHR narratives using BERT embeddings and LASSO-selected radiomic features. Their multimodal architecture achieved an AUC of 0.8449 for survival classification, substantially outperforming unimodal baselines, including a clinical/EHR-only model with AUC of 0.766. Deng et al. [35] introduced an attention-based cross-modal fusion model combining histopathological images and RNA-sequencing data. Their method improved prognostic concordance (C-index of 0.6587) relative to unimodal configurations (0.577 for image-only and 0.588 for RNA-only), highlighting the synergy of morphological and molecular inputs. Wu et al. [158] developed Lite-ProSENet, a lightweight deep learning framework that fuses CT imaging features with structured clinical data. The model achieved C-indices as high as 0.893 on the TCIA NSCLC cohort, demonstrating the benefits of combining anatomical imaging with encoded clinical variables. Hua et al. [61] focused on predicting resistance to osimertinib by integrating histopathological slides, next-generation sequencing (NGS) genomic data, demographic variables, and treatment history. Their multimodal system attained a C-index of 0.82, outperforming unimodal models (C-index range: 0.75–0.77), and illustrating the utility of comprehensive data fusion in therapeutic outcome prediction.

Despite the demonstrated promise of multimodal AI frameworks in NSCLC prognosis, several challenges remain. Existing approaches often underutilize the rich information embedded in unstructured clinical narratives, relying instead on structured variables that may miss subtle yet clinically relevant details. Moreover, unstructured text is typically treated as a sequence of tokens, without explicitly extracting task-oriented entities such as tumor stage, mutation status, or adverse events, which can limit both predictive performance and interpretability. On the imaging side, many models operate on coarse volumetric representations or precomputed radiomic features, potentially overlooking slice-level or subregional heterogeneity that is critical for survival and treatment response prediction. Fusion strategies that combine unstructured clinical text with CT imaging, while effective, are often simplistic and may fail to capture complex cross-modal interactions, limiting generalizability across diverse NSCLC cohorts. Finally, despite improvements in predictive accuracy, most multimodal models provide limited insights into the clinical rationale behind their predictions, which poses a barrier to real-world adoption.

1.5 Contributions and Thesis Outline

The aim of this thesis is to advance predictive modeling in NSCLC by leveraging unstructured clinical data from EHRs and combining it with imaging information through multimodal deep learning architectures. This work draws upon methodologies from clinical NLP, radiomics, and deep neural networks to investigate both unimodal and multimodal approaches for key prognostic tasks, including OS and pCR prediction.

The main contributions of this thesis are summarized as follows:

- **A domain-specific named entity recognition (NER) system** tailored to unstructured clinical narratives in NSCLC. This includes the construction of a task-oriented ontology of lung cancer–specific entities and the fine-tuning of a transformer-based NER model on expert-annotated datasets.
- **A novel text-based prognostic modeling framework**, where the extracted clinical entities are converted into patient-level representations using hierarchical attention networks. These embeddings are used to predict survival and treatment response outcomes, outperforming baselines based on structured variables alone.
- **A slice-level attention model for CT imaging**, which processes axial slices through 2D convolutional neural networks and aggregates spatial information via attention mechanisms. This architecture captures tumor-related patterns effectively and achieves competitive performance on public NSCLC datasets.
- **A multimodal fusion framework** that jointly models unstructured clinical text and CT imaging data. The proposed architecture integrates pretrained text embeddings and convolutional image features using attention-based fusion strategies, significantly improving survival prediction performance over unimodal counterparts.
- **Extensive experimental evaluation**, including ablation studies, baseline comparisons, and model interpretability analyses. These analyses validate the contribution of unstructured data and multimodal integration to predictive accuracy and clinical relevance.

The thesis is organized as follows. Chapter 2 focuses on the use of unstructured EHR data, with particular emphasis on free-text clinical notes, and introduces a domain-specific NER model designed to extract clinically meaningful entities relevant to NSCLC prognosis. Chapter 3 presents a deep learning framework that leverages the extracted clinical entities

Table 1.1: Datasets, modalities, endpoints, and experimental settings across thesis chapters

Chapter	**Clinical Task**	**Dataset**	**N**	**Modalities**	**Endpoint**	**Metric / Split**
2	Named Entity Recognition	Institutional NSCLC Cohort	758 docs; 257 patients	Clinical notes	25 entity types	F1-score/10-fold CV + patient split
3 (OS)	Survival prediction	Institutional NSCLC Cohort	829 docs; 231 patients	NER + structured EHR	OS	C^{td}-index/ 10-fold CV + patient split
3 (pCR)	Treatment response	Institutional NSCLC Cohort	134 docs; 67 patients	NER + clinical features	pCR	AUC/10-fold CV + patient split
4	Survival prediction	LUNG1; Institutional NSCLC Cohort	422 + 119 patients	Chest CT	2-year OS	C^{td}-index/10-fold CV
5	Multimodal mortality prediction	INSPECT dataset	19402 patients	EHR + CTPA + reports	1/6/12-mo mortality	MCC/official split

for predictive modeling, where patient-level representations are generated through hierarchical attention networks and evaluated on survival and treatment outcome prediction tasks. Chapter 4 explores predictive modeling based on chest CT imaging, proposing a novel slice-level attention model that processes individual axial slices using convolutional layers and aggregates spatial information through attention mechanisms, demonstrating strong performance on publicly available NSCLC datasets. Chapter 5 integrates the modalities addressed in the previous chapters, clinical text and CT imaging, into a multimodal fusion framework for survival prediction, evaluating early, late, and hybrid fusion strategies and showing that the combined model consistently outperforms unimodal approaches. Finally, Chapter 6 concludes the thesis by summarizing the main contributions, discussing implications for real-world clinical deployment, and outlining future research directions, including the integration of additional modalities such as genomics and longitudinal data.

1.6 Overview of Datasets and Experimental Settings

To provide a clear and unified overview of the experimental framework adopted throughout this thesis, Table 1.1 summarizes the datasets, cohort sizes, clinical endpoints, modalities, and data splitting strategies used across the chapters. This overview aims to facilitate cross-chapter comparison and improve the overall readability and reproducibility.

Chapter 2

Information Extraction from Unstructured EHR

The growing availability of EHRs has ushered in new opportunities for improving the understanding, diagnosis, and treatment of complex diseases such as NSCLC. Although structured data in EHRs is organized and standardized patient information in fixed fields, such as diagnostic codes and lab results, a significant portion of clinically valuable information resides in unstructured narratives. These include physicians' notes, imaging reports, and follow-up summaries, which often contain longitudinal insights, personalized treatment details, and contextual nuances not captured elsewhere.

Extracting structured information from these unstructured texts is a critical step toward enabling data-driven decision-making in oncology. Named Entity Recognition (NER), a core task in NLP, facilitates this transformation by automatically identifying and classifying mentions of medical concepts within free-text documents. In recent years, the application of deep learning and transformer-based models has substantially improved the performance of NER systems, particularly in English-language biomedical texts. However, limited research has addressed the challenges of clinical NER in less-resourced languages such as Italian, especially in the context of NSCLC.

This chapter explores the role of NER in structuring unstructured EHR data, with a focus on the development and evaluation of a domain-adapted NER system for Italian-language clinical reports of NSCLC patients. We present the conceptual, methodological, and architectural components of the system, detailing the ontology of clinical entities, the annotation framework, and the neural modeling techniques used. Finally, we benchmark the proposed approach against general-purpose language models, highlighting the benefits of domain-specific pretraining.

2.1 Motivation and Clinical Background

The complexity of NSCLC diagnosis, treatment, and monitoring requires the systematic recording of a wide array of clinical information, which is often stored in EHRs. Although structured data formats such as ICD codes or laboratory test results provide standardized insights, much of the rich and nuanced information describing patient condition, treatment rationale, and disease progression is embedded in unstructured clinical narratives, including oncology consultations and radiotherapy notes.

These unstructured texts represent a valuable source for understanding longitudinal patient history, symptom evolution, and real-world treatment decisions. However, mining actionable information from such narratives is challenging due to their free-text format, domain-specific terminology, and variable syntax. NLP techniques, particularly Named Entity Recognition (NER), provide a scalable approach for extracting and categorizing clinically relevant information from unstructured content (Figure 2.1). NER aims to identify and categorize spans of text corresponding to clinical concepts, such as symptoms, diagnoses, therapies, laboratory findings, and procedures. Importantly, the extracted entities are not necessarily normalized or codified into canonical, discrete variables, but often consist of multi-token expressions that preserve the original linguistic formulation. As such, the output of NER in this work should be interpreted as a semi-structured representation of clinical information, in which relevance is captured through contextualized textual embeddings rather than explicit numerical encoding.

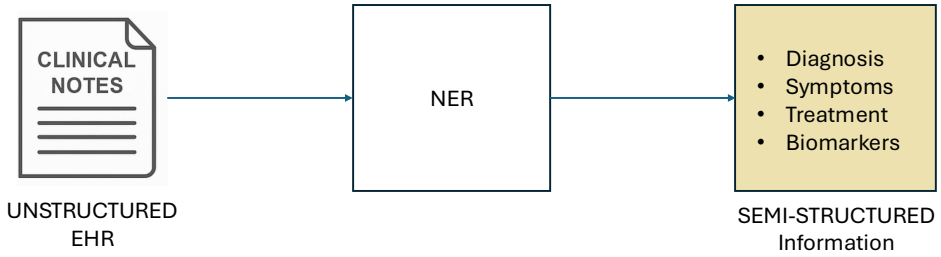

Figure 2.1: Overview of the information extraction process using NER in lung cancer care. Unstructured EHR data, such as clinical notes and radiology reports, are processed by an NER system to identify and categorize clinically relevant concepts, including diagnoses, symptoms, treatments, and biomarkers, yielding semi-structured, embedding-based representations.

This semi-structured extraction paradigm enables the reuse of information embedded in clinical narratives without imposing strong assumptions on canonical coding or normalization. In the context of lung cancer care, such representations can be directly leveraged in

downstream predictive models, such as OS, by operating on entity-level embeddings rather than on manually engineered or fully structured feature vectors.

Nevertheless, the development of robust NER systems for healthcare remains a non-trivial task. Clinical language is characterized by ambiguous abbreviations, institution-specific jargon, and a lack of grammatical regularity, which substantially limits the effectiveness of naive text-processing approaches. Early NER methods primarily relied on rule-based systems or traditional machine learning classifiers, which required extensive manual feature engineering and showed limited generalization across datasets and clinical settings. Recent advances in deep learning, and particularly the adoption of transformer-based architectures such as BERT, have significantly reshaped the field. By modeling contextual dependencies and long-range semantic relationships, these models are able to better capture the nuances of clinical language, making them especially well suited for biomedical and clinical NER tasks.

To the best of our knowledge there is few work in applying BERT-based models for NER to the specific field of NSCLC and the majority includes the lung cancer patients' clinical reports in wider datasets, with the aim of extracting general oncological information or rather information about the diagnosis [60, 105]. Only two contributions tackle the NER task with the aim of extracting specific information about lung cancer patients from their clinical reports [165, 138]. The article by Zhang et al. [165] extracts entities from Chinese CT reports for lung cancer screening and TNM staging using a BERT-based Bidirectional Long-Short Term Memory (BiLSTM)-Transformer network. The network was pre-trained on 170 chinese unannotated CT reports, to address the challenge posed by limited labeled data. To evaluate the effectiveness of their approach, the researchers conducted experiments on a private clinical dataset consisting of 359 CT reports obtained from the Department of Thoracic Surgery II at Peking University Cancer Hospital, in which they identified 14 entity types about lung cancer screening and staging. To train and evaluate the proposed model, they randomly separated 70% CT reports as the training set, 10% as the validation set, and 20% as the test set. The proposed deep learning method demonstrated high efficiency in recognizing the 14 clinical entities pertinent to lung cancer screening and staging, achieving an F1-score equal to 85.96%. In [138] the authors proposed an innovative deep learning-based approach for the extraction of lung cancer information from clinical notes written in Spanish. To tackle this challenging task, they leveraged the power of two separate deep learning models: BiLSTM and pre-trained BERT. In order to evaluate the effectiveness of each approach, the researchers manually curated a private annotated corpus comprising 500 clinical notes of patients diagnosed with lung cancer where they found 19 biomedical entity types about lung cancer staging, diagnosis, treatments, patient information and lifestyle. Using a 10-fold cross-validation strategy results showed an F1-score of 91.08% and 93.72% for BiLSTM and

BERT models, respectively. Although both [165] and [138] achieve impressive performance in accurately identifying and extracting relevant information related to lung cancer from the clinical reports, a major limitation of the literature on NER for lung cancer is the adaptation of transformer-based models for the biomedical domain in less-resourced languages as Italian. Indeed, the majority of existing research in this field has predominantly focused on the English language, thanks to the availability of large public datasets such as MIMIC [70], and more recently on other languages, as Chinese [165] and Spanish [138]. The task of extracting named entities from oncology clinical texts written in Italian is largely unexplored, needing domain-specific checkpoints and access to public biomedical data. A recent attempt to start filling this gap in the Italian biomedical field is the pre-trained model called Biomedical BERT for Italian (BioBIT) [15]. BioBIT aims to address the disparity between Italian and English NLP by employing Italian machine translations of English resources and by integrating a domain specific corpus composed of Italian texts. The pre-training of this Italian biomedical checkpoint can be summarized as follows. Initially, a general-purpose Italian BERT checkpoint called BaseBIT [15] was pre-trained on a non-medical Italian corpus of 81 GB and 13 billion tokens. This checkpoint was further augmented with a biomedical corpus obtained by automatically translating the MIMIC [70] corpus, comprising over 2 million PubMed abstracts for a final dimension of 28 GB. The resulting checkpoint was called BioBIT (Biomedical BERT for ITalian). To increase the quality of BioBIT pre-training, the authors of [15] developed two additional checkpoints (MedBITR3+ and MedBITR12+) by further pre-training BioBIT on a 100 MB private corpus of medical textbooks used in formal medical doctors' education and specialized training, which were either directly written by Italian authors or translated by human professional translators.

The three checkpoints were evaluated on 6 NER tasks using different public datasets translated from English to Italian, targeting the annotation of chemicals, diseases, genes, and other biomedical-related mentions. The MedBITR3+ and MedBITR12+models were able to outperform BioBIT checkpoint on 4 over 6 NER tasks, proving the benefits of the higher quality training data. Despite the fact that BioBIT checkpoints represent a significant step forward in the field of biomedical NER in Italian, it is worth noticing that all datasets used to test the network on NER tasks were translated from English to Italian, hence the proposed checkpoints were not evaluated on an original Italian dataset, or rather on a real-world NER task specifically focusing on lung cancer in the Italian language. For these reasons, we propose this work that represents a first attempt to apply NER on a real-world Italian dataset on NSCLC.

Table 2.1: Entity types acronyms and descriptions

Entity type	Acronym	Description
Cancer	CAN	Physicians' descriptions of tumors (e.g. 'adenocarcinoma') and metastasis concepts (e.g. 'bone metastasis').
Comorbidity	COM	Diseases or conditions that co-occur with a cancer diagnosis.
Cancer stage	STA	The stage of the tumor at the time of diagnosis. It may be implicitly described (e.g. 'advanced stage') or indicated using a scale ranging from I to IV.
Focal anomaly	FAN	Any type of abnormality or suspicious pathology, such as nodules and lesions.
Anatomical position	POS	The specific anatomical location of the cancer or anomaly, such as the right lung.
Therapy	TPY	The name of the therapy used to treat patients, including radiotherapy and surgery.
Drug	DRU	The names of drugs used in the treatment of cancer patients.
Dosage	DOS	The dosage of drugs (e.g., 25 mg) and therapy.
Therapy duration	DUR	The duration of a patient's cancer treatment or the period during which a specific drug was administered.
Teraphy line	TPL	The number of cycles within a therapy.
Tumor progression	TUP	Changes in the rate of growth or invasiveness of cancer cells.
Quantity habits	QHA	The quantity of cigarettes smoked by the patient or alcohol consumed.
Exam	EXA	All the medical examinations undergone by a cancer patient.
Medication frequency	FRE	The frequency of medication administration.
Patient event	PEV	When a treatment has been given to a patient, reduced, changed, or discontinued.
Patient symptom	PSY	Symptoms experienced by the patient.
Mass	MAS	Abnormal growth of cells that forms a mass or tumor within the tissue. It can be described in terms of maximum diameter (mm or cm) or volume (cm^3)
Morphology	MOR	Shape and structure of the tumor, such as a solid formation with irregular margins.
Date	DAT	Dates of exams, diagnosis, and follow-up implicity mentioned in the clinical notes.
Histology	HIS	Histological characteristics of the cancer, such as 'squamous'.
Familiarity	FAM	Cancer cases in the patient's family clinical history.
TNM classification	TNM	T describes the tumor's size on a scale of 1 to 4, N indicates the status of nearby lymph nodes (ranging from 0 to 3), and M indicates the presence of distant metastasis (0 if absent, 1 if present).
Numerical Rating Scale	NRS	Pain level on a scale from 0 to 10, with 0 indicating no pain and 10 representing the worst tolerable pain.
Weight	WEI	Weight of a cancer patient.
Height	HEI	Height of a cancer patient.

2.2 Materials and Methods

2.2.1 Clinical Concept Taxonomy and Annotation Framework

A fundamental step in developing an effective NER system is the definition of a domain-specific ontology that accurately captures the clinical concepts of interest. In the context of NSCLC, a set of 25 clinically meaningful entity types, listed in Table 2.1, was defined in collaboration with an oncologist and a radiation oncologist. These entities cover a broad range of information relevant to lung cancer management, including cancer characteristics (e.g.,

morphology, stage, TNM classification), patient-related factors (e.g., comorbidities, symptoms, familiarity), treatment details (e.g., drug, dosage, therapy duration), and diagnostic information (e.g., imaging findings, histology).

This ontology was tailored to address the specific needs of NSCLC care and ensures that the resulting annotations support both clinical interpretability and computational feasibility. The decision to limit the entity set to those requiring genuine NER, rather than simpler extraction techniques, further enhances the precision and relevance of the annotation effort.

The annotation process was carried out using the Doccano tool [109] and involved the manual labeling of 758 Italian clinical reports from 257 patients diagnosed with stage III or IV NSCLC [127, 19]. These reports, collected prior to the initiation of therapy, are divided into two main categories: oncological visits and radiotherapy visits. Each document provides a comprehensive overview of the patient's clinical profile, including personal information, medical history, reason for consultation, histological and imaging findings, physical examinations, preliminary diagnoses, prescriptions and recommendations, conclusions, and follow-up plans.

The patient cohort was enrolled under two separate ethical approvals: the first approved on 30 October 2012 and registered at ClinicalTrials.gov on 12 July 2018 (Identifier: NCT03583723); the second approved on 16 April 2019 (Identifier: 16/19 OSS).

Annotations were performed using the BIO (Beginning, Inside, Outside) tagging format to mark the spans corresponding to each entity type. For instance, the sentence 'Adenocarcinoma dell'apice polmonare sinistro' (Adenocarcinoma of the left lung apex) can be labelled as shown in Figure 2.2. It is worth noticing that in this example an entity can include one or more tokens: the entity type CAN is assigned to an entity encompassing one token (i.e. 'Adenocarcinoma'), whereas the entity type POS is assigned to an entity encompassing three tokens (i.e. 'apice', 'polmonare', 'sinistro').

Figure 2.2: Sentence labeled with the BIO tagging format.

Initial annotation was carried out by a trained non-expert and subsequently validated by clinicians to ensure clinical soundness.

To evaluate the quality and consistency of the annotations, inter-annotator agreement (IAA) was measured using the F1-score. This choice is motivated by findings from several studies [17, 133, 76], which argue that the F1-score [59] is more appropriate than the Kappa statistic [29] in the context of named entity annotation. This is primarily because annotated entities can span varying numbers of tokens, making token-level agreement metrics like Kappa less reliable. On the one hand, we first computed the score at token level, meaning that the IAA for each token is calculated separately. Indeed, when an entity contains several tokens, a pair of annotators can annotate different token groups. In this case, correct annotations include only those tokens where both annotators agree. On the other hand, the IAA F1-score was also obtained at the entity level, meaning that the correct annotations encompass only those entities where both annotators agree on all annotated tokens.

Table 2.2: Inter-annotator agreement for the NSCLC corpus

Entity type	**Token level IAA**	**Entity level IAA**
Cancer	0.92	0.92
Comorbidity	0.95	0.95
Cancer stage	1.0	1.0
Focal anomaly	0.97	0.97
Anatomical position	0.98	0.98
Therapy	0.99	0.99
Drug	0.99	0.98
Dosage	1.0	1.0
Therapy duration	1.0	1.0
Teraphy line	1.0	1.0
Tumor progression	1.0	1.0
Quantity habits	1.0	1.0
Exam	0.99	0.99
Medication frequency	1.0	1.0
Patient event	0.98	0.98
Patient symptom	0.93	0.96
Mass	1.0	1.0
Morphology	1.0	1.0
Date	1.0	1.0
Histology	1.0	1.0
Familiarity	1.0	1.0
TNM classification	1.0	1.0
NRS	0.82	0.57
Weight	1.0	1.0
Height	1.0	1.0
Average	0.98±0.04	0.97±0.08

Table 2.2 shows the IAA using the F1-score: on the one hand with respect to the token level IAA values are near the complete agreement for all entities, on the other hand looking at the entity level IAA we notice a lower value for the NRS entity type. We deem this result depends on the difficulty in agreeing with the correct number of tokens comprising

this entity, which is one of the main challenges when dealing with a NER annotation process. However, looking at the average values for both measures, it is possible to conclude that the annotation is overall reliable.

2.2.2 Neural Architecture for Clinical NER

The technical implementation of a clinical NER system relies on the careful design of the underlying neural architecture. A transformer-based model was employed, specifically the MedBITR3+ checkpoint, a variant of BioBIT pretrained on Italian biomedical corpora [15]. This model integrates both general linguistic knowledge and domain-specific information, rendering it well-suited for extracting entities from clinical narratives in Italian.

As depicted in Figure 2.3, the training pipeline followed a three-stage process: (1) corpus generation and preprocessing, (2) fine-tuning of the pretrained model, and (3) evaluation using cross-validation. Preprocessing included sentence segmentation and tokenization adapted for clinical Italian text of annotated NSCLC clinical reports. To address linguistic and structural challenges of NSCLC clinical reports written in Italian, we developed a custom algorithm for sentence detection and tokenization specifically designed for the Italian NSCLC clinical notes. First, the reports in the annotated corpus were split into distinct sentences using two separator characters: the dot character '.' and the double new line character '/n/n'. The first character ('.') was considered a separator, except when: it appeared after abbreviations; it was followed by a number; in special cases like web addresses and emails.

The second separator ('/n/n'), was chosen since typically two occurrences of '/n' character indicate the start of a new sentence, unlike a single occurrence. After splitting the documents into sentences, each sentence was tokenized, i.e. split in its atomic units using various word separators, including the blank space, brackets, punctuation marks, etc..

The fine-tuning phase optimized the model on the annotated NSCLC corpus for the specific NER task addressing class imbalance via focal loss. Finally, a rigorous 10-fold cross-validation strategy ensured robust evaluation.

Among the transformer-based architectures available for the Italian language, the MedBITR3+ checkpoint [15] has been demonstrated to be particularly effective for biomedical applications. This model builds upon the BaseBIT [15] and BioBIT [15] checkpoints by incorporating additional pretraining on high-quality, domain-specific data, namely, Italian medical textbooks written or translated by experts.

This domain adaptation strategy allows the model to acquire an in-depth understanding of specialized vocabulary, syntactic patterns, and semantic nuances characteristic of clini-

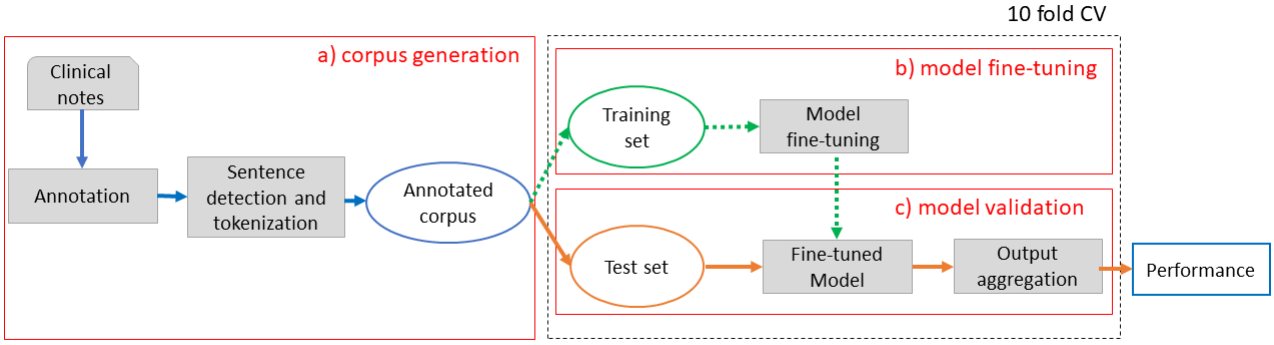

Figure 2.3: Proposed pipeline. Panel a) shows the corpus generation, including annotation and the pre-processing of the raw text (sentence detection and tokenization). Panel b) shows the fine-tuning phase, whereas panel c) the validation phase. Both b) and c) are carried out considering a 10 fold cross-validation experimental setup (10 fold CV black dotted box).

cal discourse. Compared to general-purpose models such as multilingual BERT (mBERT) [36] and umBERTo [119], which are trained on Wikipedia or CommonCrawl corpora, MedBITR3+ provides a distinct advantage in accurately identifying entities specific to oncology and NSCLC care.

2.3 Experimental Setup

The model fine-tuning was performed on the annotated NSCLC dataset using the following configuration: a batch size of 8, 12 training epochs, a dropout rate of 0.1, and a learning rate of 5×10^5 with the Adam optimizer. These settings were selected to balance convergence stability with generalization performance.

A focal loss function [96], defined as

$$\mathbf{FL}(p_t) = \text{-}\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{2.1}$$

was employed to mitigate the effects of class imbalance inherent in the dataset, which features highly skewed distributions of certain entity types (Figure5.4). The loss function dynamically down-weights well-classified examples and emphasizes harder, minority-class samples. This approach enables the model to better learn infrequent but clinically significant entities.

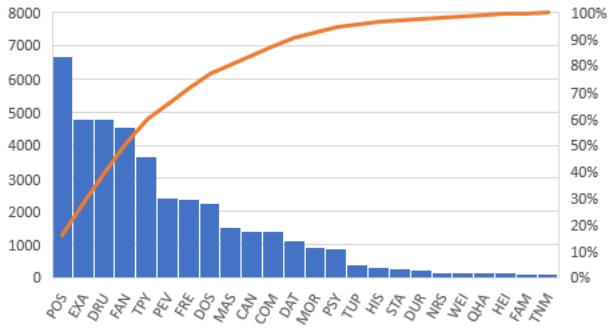

Figure 2.4: Histogram of entity types. On the y-axis we show the count (on the left) and the percentage frequency (on the right) of occurrences of each entity type. On the x-axis we show the various entity types. In addition to the histogram, we also display the Lorenz curve (in orange), which illustrates the distribution of entities in terms of their occurrences.

2.4 Results and Discussion

Evaluating the performance of a NER system within a clinical context necessitates a rigorous and comprehensive assessment strategy. To this end, the MedBITR3+ model [15] was benchmarked against two general-purpose transformer-based models, multilingual BERT (mBERT) [36] and umBERTo [119], using standard evaluation metrics: precision (P), recall (R), and F1-score (F1). These metrics were computed at the entity level under a strict matching criterion, which requires complete span overlap between predicted and annotated entities.

$$P(E_t) = \frac{\#\text{Entities correctly predicted as } E_t}{\#\text{Entities predicted as } E_t} \tag{2.2}$$

$$R(E_t) = \frac{\#\text{Entities correctly predicted as } E_t}{\#\text{Entities belonging to } E_t \text{ in the dataset}} \tag{2.3}$$

$$F1(E_t) = \frac{2 \cdot P(E_t) \cdot R(E_t)}{P(E_t) + R(E_t)} \tag{2.4}$$

Equations 2.2, 2.3 and 2.4 present the formal computation of the aforementioned metrics according to this rationale. Each score is calculated with respect to a specific entity type E_t and the symbol '#' stands for 'Number of'.

To ensure robust and unbiased evaluation, a 10-fold cross-validation was employed at the patient level. This approach minimizes the risk of overestimating performance by preventing the leakage of patient-specific information between training and test sets. The predictions from each fold were aggregated to produce overall performance scores across the entire dataset.

Table 2.3: Results for multilingual BERT, umBERTo, MedBITR3+

	mBERT			umBERTo			MedBITR3+		
Entity type	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
Cancer	0.927	0.906	0.949	0.934	0.914	0.955	**0.949**	**0.934**	**0.965**
Comorbidity	0.679	0.664	0.694	0.733	0.702	0.767	**0.762**	**0.744**	**0.780**
Cancer stage	0.791	0.750	0.837	0.786	0.753	0.822	**0.800**	**0.761**	**0.843**
Focal anomaly	0.800	0.770	0.834	0.802	0.769	0.837	**0.818**	**0.789**	**0.849**
Anatomical position	0.770	0.747	0.794	0.783	0.754	0.815	**0.801**	**0.779**	**0.824**
Therapy	0.844	0.818	0.873	0.847	0.814	0.882	**0.865**	**0.840**	**0.891**
Drug	0.899	0.893	0.905	**0.921**	**0.915**	0.926	**0.921**	0.911	**0.931**
Dosage	0.897	0.886	0.909	0.908	0.901	0.915	**0.917**	**0.914**	**0.920**
Therapy duration	0.761	0.727	0.798	**0.785**	**0.757**	0.814	0.764	0.716	**0.818**
Teraphy line	0.877	0.875	0.879	**0.894**	**0.865**	**0.925**	0.892	0.861	**0.925**
Tumor progression	0.517	0.484	0.55	0.603	0.565	0.645	**0.623**	**0.580**	**0.672**
Quantity habits	**0.911**	**0.911**	0.911	0.898	0.876	**0.923**	**0.911**	0.906	0.917
Exam	0.870	0.850	0.891	0.874	0.851	0.898	**0.892**	**0.872**	**0.914**
Medication frequency	0.797	0.758	0.842	0.807	0.764	**0.855**	**0.815**	**0.779**	0.854
Patient event	0.757	0.717	0.802	0.754	0.703	0.814	**0.778**	**0.735**	**0.826**
Patient symptom	0.683	0.668	0.699	0.748	0.720	0.778	**0.766**	**0.751**	**0.781**
Mass	0.872	0.841	0.906	0.879	0.851	0.910	**0.895**	**0.875**	**0.915**
Morphology	0.693	0.653	0.739	0.712	0.674	0.754	**0.740**	**0.703**	**0.782**
Date	0.905	0.871	0.940	0.904	0.871	0.940	**0.915**	**0.883**	**0.949**
Histology	0.885	0.880	0.891	**0.917**	**0.908**	0.925	0.907	0.885	**0.931**
Familiarity	0.601	0.544	0.684	**0.646**	0.587	**0.718**	0.643	**0.594**	0.701
TNM classification	0.776	0.693	0.884	**0.795**	**0.691**	**0.938**	0.780	0.678	0.919
NRS	0.980	0.967	0.994	0.959	0.931	0.989	**0.969**	**0.946**	**0.994**
Weight	0.938	0.928	0.949	0.958	0.939	**0.978**	**0.971**	**0.977**	0.966
Height	**0.976**	0.969	**0.981**	0.942	0.939	0.945	0.969	**0.981**	0.957
Average ± std	0.817±0.113	0.791±0.124	0.846±0.103	0.832±0.09	0.800±0.110	0.867±0.085	**0.843±0.094**	**0.816±0.109**	**0.873±0.082**

2.4.1 Benchmarking Against General-Purpose Language Models

The benchmarking analysis aimed to quantify the benefit of employing a domain-specific language model by comparing the MedBITR3+ checkpoint against two widely adopted general-purpose models: multilingual BERT (mBERT) and umBERTo. While mBERT is trained on Wikipedia across 104 languages, including Italian, umBERTo is trained specifically on Italian text from the OSCAR corpus. However, neither model is pretrained on biomedical data, which limits their effectiveness in clinical NER tasks.

The upper panel of Table 2.3 shows results of mBERT, UmBERTo and MedBITR3+ for each entity type, whereas the lower panel presents the average performance with the standard deviation (std). The highest scores are highlighted in bold. Looking at the results it is evident that the MedBITR3+ model outperforms the others. The MedBITR3+ model achieved an average F1-score of 0.843, outperforming mBERT and umBERTo, which reached 0.817 and 0.832, respectively. Notably, MedBITR3+ attained the highest F1-score across 76% of the 25 entity types, demonstrating its superior ability to generalize across diverse clinical concepts. For instance, in the identification of cancer-related terms (e.g., morphology, staging, and drug therapy), the domain-adapted model consistently delivered higher precision and recall.

In addition to descriptive statistics, the McNemar statistical test [82] was conducted to assess the significance of the performance differences between MedBITR3+ and its competitors. In both comparisons the final McNemar statistic s was far greater than the threshold value 3.841 ($s = 6682.90$ for MedBITR3+ vs. umBERTo and $s = 10527.90$ for MedBITR3+ vs. mBERT) with a p-value near 0. These results demonstrate the statistically significant superiority of MedBITR3+ in the NER task at hand, thereby confirming that the observed gains were not due to random variation.

In conclusion, the benchmarking results affirm the added value of domain-adaptive pre-training in biomedical NLP. Models like MedBITR3+, which integrate specialized clinical language into their training corpus, are better equipped to handle the intricacies of healthcare narratives and deliver more reliable performance in information extraction tasks relevant to lung cancer care.

Chapter 3

Prognostic Modeling from NER-Driven Representations of Unstructured Clinical Text

The narrative sections of EHRs are a rich repository of patient information, encompassing diagnostic findings, treatment plans, comorbidities, and disease progression notes. In the case of NSCLC, such unstructured text can contain subtle prognostic cues that may not be reflected in structured data fields.

However, the inherent variability of clinical language, use of abbreviations, and lack of consistent formatting present major obstacles to direct computational analysis.

NLP, and in particular NER, provides a mechanism to transform these free-text narratives into structured, semantically meaningful representations. Although much of the existing literature has focused on using NER for feature extraction, relatively little attention has been given to leveraging the contextual embeddings produced by NER models themselves. These embeddings preserve not only the presence of medical entities but also the linguistic and clinical context in which they appear, potentially carrying latent prognostic information that manual extraction might overlook.

This chapter presents a prognostic modeling framework that capitalizes on these rich NER-derived representations through a Hierarchical Embedding Attention architecture (HEAL). The approach integrates entity-level contextual embeddings from a transformer-based multiclass NER system with a two-level attention mechanism, enabling selective aggregation of information first at the token level and then across sentences and reports. The resulting patient representation serves as input to predictive models for two clinically relevant endpoints:

- **OS**, a time-to-event outcome measuring the duration from diagnosis or treatment to death.
- **pCR**, a binary endpoint indicating the absence of residual invasive cancer after neoadjuvant treatment.

Beyond predictive performance, the proposed methodology emphasizes interpretability, providing attention maps that highlight clinically relevant sentences and entities. By comparing this automated, data-driven representation with manually curated clinical features, we assess the potential of NER-driven embeddings to enhance both the accuracy and transparency of prognostic models.

The remainder of this chapter details the motivation for structured semantics from free text (Section 3.1), the hierarchical representation learning architecture (Section 3.2), the clinical outcomes addressed (Section 3.3), the experimental design (Section 3.4), quantitative evaluation (Section 3.5), and interpretability analysis (Section 3.6).

3.1 Motivation and Clinical Background

EHRs often contain rich narrative sections that capture nuanced clinical details unavailable in structured tables. In the context of NSCLC prognosis, these free-text notes include histology descriptions, comorbidities, treatment regimens, and temporal disease evolution, signals that are critical for accurate survival prediction. However, their unstructured nature poses challenges for computational analysis. As seen in the previous chapter, NER offers a means of converting these narratives into structured semantic units, identifying and categorizing clinically relevant concepts such as cancer stage, therapy, comorbidities, and histology (Figure 3.1).

Early progress in clinical NER was slower than in the general domain, largely due to limited availability of annotated clinical data. Over time, the release of public benchmarks and shared tasks, together with the adoption of machine learning and deep learning approaches such as CRFs [83] and BiLSTMs [56], contributed to advancing the field, as discussed in Chapter 2. More recently, transformer-based pre-trained language models built on the BERT architecture have emerged as the dominant paradigm for clinical NER, consistently outperforming earlier methods across multiple benchmarks [111]. An overview of representative recent approaches applied to EHR data is reported in Table 3.1. Specifically, we observe that 8 out of 12 studies employ a BERT-based approach, which aligns with the methodology adopted in this work. With respect to the clinical domain, most of the analyzed papers address broad and heterogeneous clinical scenarios, while only three focus on a

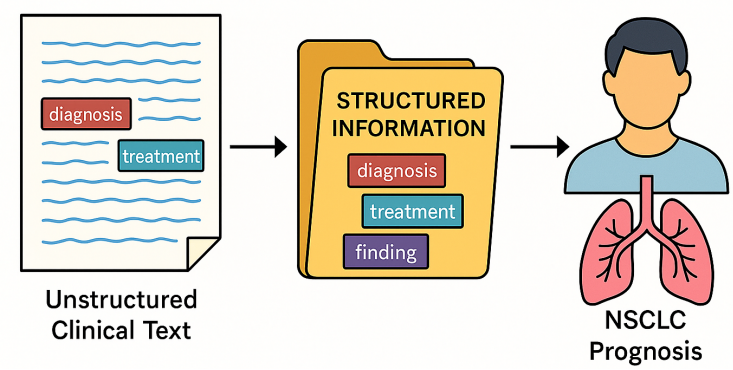

Figure 3.1: Conceptual overview of transforming unstructured clinical text into structured, entity-level information for prognostic modeling in NSCLC. The process begins with raw narrative EHR data, applies NER to extract clinically relevant entities (e.g., diagnosis, treatment, findings), and generates structured representations that inform predictive models for patient survival outcomes.

specific pathology [174, 166, 139]. This predominance of general-purpose settings limits the ability to capture disease-specific linguistic patterns, clinical terminology, and contextual nuances that may be critical for modeling complex conditions such as lung cancer. As a result, clinically relevant aspects, including pathology-specific symptoms, treatment protocols, and prognostic descriptors, may be underrepresented or insufficiently characterized in existing approaches. Additionally, nearly all of these studies formulate NER as a multiclass problem, reflecting the simultaneous presence of multiple entity types within clinical narratives. In light of this, our approach involves implementing a multiclass NER system. We believe that the distinctive classes within the NER embeddings can provide significant benefits when using such representations to train a predictive model. Among the literature examined, we found only three papers that include NER to build a predictive model, but they do not use NER embeddings as feature representations. Instead, entities are extracted through a NER system and subsequently transformed into numerical representations, primarily using various embedding techniques such as BERT-based models or Word2Vec. This process, however, results in the loss of EHRs contextual information surrounding the entities, which can provide valuable insights for a more comprehensive and accurate understanding. By representing entity-level information in a semi-structured form through contextual embeddings, we can create richer and more discriminative patient representations that capture both the identity of clinical entities and aspects of their surrounding context. These representations bridge the gap between raw narrative text and survival modeling, allowing predictive models to leverage detailed, clinically relevant information while remaining automated, scalable, and interpretable.

Table 3.1: Recent advances in NER applied to EHR. Abbreviations: BP=Body Part, Trt=Treatment, Sx=Symptom, MedProb=Medical Problem, Op=Operation, Med=Medicine, Lab=Laboratory, Exam=Examination, Dos=Dosage, Freq=Frequency, RoA=Route of Administration, Str=Strength.

Model	**Ref.**	**Dataset**	**App.**	**Entities**	**Perf.**	**Usage**
MC-BERT+ BiLSTM+ CNN+ MHA+ CRF	[23]	CCKS17 [90], CCKS19 [176], cEHRNER [167]	Notes	9: BP, Trt, Signs, Check, Disease, Lab, Med, Op, Sx	F1: 94.2, 86.5, 92.3	-
BiLSTM-CNN-Char	[78]	i2b2-10 [154], 2014 n2c2 [142], 2018 n2c2 [54]	Notes	4: MedProb, Trt, Test, Drug	F1: 87.6, 96.1, 89.9	-
MUSA-BiLSTM-CRF	[8]	CCKS17 [90], CCKS18 [170]	Notes	5: Disease, Sx, Exam, Trt, BP	F1: 92.0, 91.8	-
BERT	[110]	2018 n2c2 [54], 2009 n2c2 [123], 2010 n2c2 [154], 2012 n2c2 [143], ShARe13 [43]	Notes	4: Drugs, Dos, Reasons, ADE	F1: 90.0, 80.9, 88.4, 87.5, 82.6	-
BERT-BiLSTM-CRF	[74]	ShARe13 [43], ShARe14 [18]	Notes	1: Disorder	F1: 79.9, 80.7	-
BERT	[124]	i2b2-2010 [154], VietBioNER [125]	Notes	3: MedProb, Trt, Test	F1: 87.7, 80.9	-
CancerBERT	[174]	Proprietary (EHRs)	Breast cancer NER	8: HR type, HR status, Tumor size, site, grade, type, laterality, stage	F1: 87.6	-
scispaCy	[104]	MIMIC-III [70]	Notes	2: Disease, Chemical	-	Mortality
med7	[169]	MIMIC-III [70]	Notes	7: Dos, Drug, Duration, Form, Freq, RoA, Str	-	Mortality
Rule-based	[166]	CCKS20 [91], gastroscopy, mixed	Breast cancer NER	6: Disease, Anatomy, Imaging, Lab, Drug, Op	F1: 87.9, 99.8, 96.2	-
Ensemble (CRF, mBERT, XLM-RoBERTa, LSTM)	[71]	Proprietary (EHRs)	Notes	11: Dept, Date, Duration, Evidential, Freq, Occurrence, Problem, Test, Time, Trt, Value	F1: 89.2	-
RoBERTa	[139]	Proprietary (EHRs)	Breast cancer info	23: Breast cancer domain	F1: 95.0	-

3.2 Materials and Methods

3.2.1 Datasets and Preprocessing

Two datasets were used:

- **OS prediction**: 829 pre-treatment clinical reports from 231 stage III–IV NSCLC patients.
- **pCR prediction**: A separate cohort of 134 EHRs belonging to 67 patients with stage III NSCLC, of whom 15 had a positive pCR and 52 had a negative pCR.

The datasets were annotated for the 25 entity types described in Chapter 2. Entity annotation achieved >0.97 F1 inter-annotator agreement for both datasets. NER was performed using the best-performing model identified in Chapter 2, namely a fine-tuned biomedical BERT architecture (MedBITR3+ for Italian) [15], trained with Focal Loss to address class imbalance. Both tasks were evaluated using stratified 10-fold cross-validation at the patient level.

3.2.2 Representation Learning via Hierarchical Attention Networks

The proposed methodology, HEAL, depicted in Figure 3.2, transforms NER-derived entity embeddings into patient-level representations through a hierarchical, two-level attention mechanism. Starting from a transformer-based multiclass NER model fine-tuned on NSCLC-specific entity types, the architecture first aggregates embeddings of entity tokens at the sentence level, and then across all clinical notes for a given patient, producing a rich patient representation that encodes both entity identity and contextual information.

This patient representation is then fed into a prediction network. For OS prediction, the network functions as a multi-output risk assessment model, providing hazard estimates at the considered time points. For pCR classification, the same representation is used as input to a separate binary classification network. In both cases, the networks leverage the semi-structured, embedding-based patient representations to enable automated, scalable, and interpretable prognostic modeling.

Token-Level Attention

For each sentence containing at least one recognized entity, the model extracts token embeddings $\mathbf{e}_j$ of size d_E from the NER encoder before classification. A soft attention mechanism assigns weights to entity tokens, reflecting their relative importance in the sentence for the

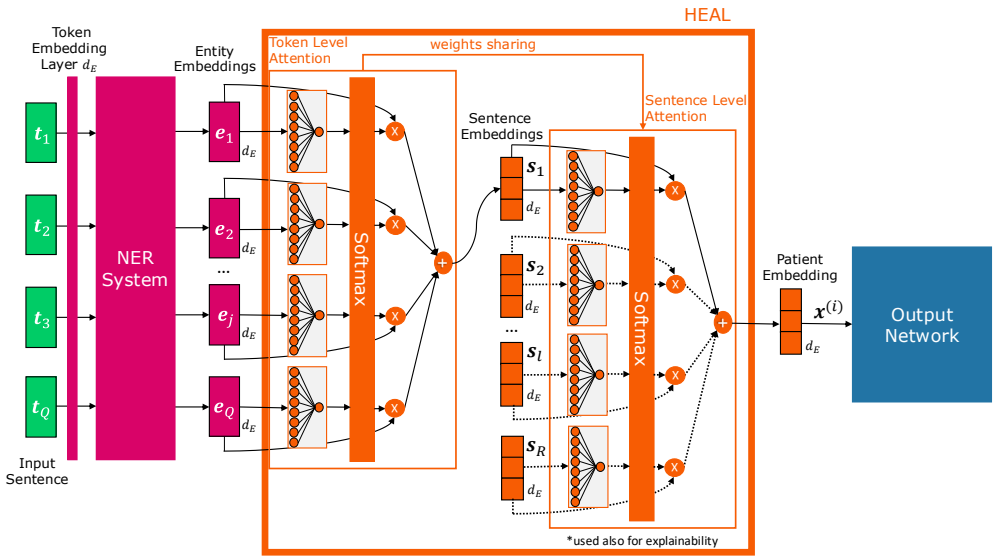

Figure 3.2: Proposed architecture: The architecture utilizes token embeddings generated by the NER system before the classification layer. Each token embedding classified by the NER system as an entity (Entity Embedding) undergoes a weighting process through a token attentional layer. This produces a weighted average of the same embedding size, named as Sentence Embedding. The sentence embeddings derived from all sentences in a patient's clinical reports, are then fed into a sentence attentional layer, which shares weights with the token attentional layer. The outcome is a weighted average vector, maintaining the original embedding size dE , referred to as the patient embedding x(i). The patient embedding is the input of the output network.

prognostic task. The weighted average of these embeddings produces a sentence embedding of fixed dimensionality $\mathbf{s}_l$ as follows:

$$\mathbf{s}_l = \sum_{j=0}^{Q} w_j \mathbf{e}_j. \tag{3.1}$$

where w_j is the weight produced by the soft attention for the token embedding $\mathbf{e}_j$, Q is the total number of tokens in the input sentence and l ranges from 1 to the total number of sentences R in the patient clinical reports that contain at least one token classified as an entity. It is worth nothing that $\mathbf{s}_l$ has the same size d_E of the token embedding. This process prioritizes clinically relevant terms, such as comorbidity mentions or histological subtypes, over less informative text.

Sentence-Level Attention

The sentence embeddings, originating from all sentences within a patient's clinical reports, are then inputted into a sentence attentional layer, i.e., the right most block in Figure 3.2 that compute

$$\mathbf{x}^{(i)} = \sum_{l=0}^{R} w_l \mathbf{s}_l. \tag{3.2}$$

It is a comprehensive representation of the patient's clinical information and serves as the vector of features for the risk assessment network in Figure 3.2. It is worth noting that this layer incorporates a soft attention mechanism that shares weights with the token attentional layer. The purpose of weight sharing is twofold: first, to decrease the overall number of network parameters, thereby mitigating overfitting on small data and reducing computational complexity; second, to enhance the transferability of features between layers since the clinical relevance of a sentence is correlated to the clinical relevance of a recognized clinical entity. Hence, the result is a weighted average vector, maintaining the same embedding size d_E and named as the patient embedding.

3.2.3 Clinical Outcomes and Prediction Network

Time-to-Event Prognostic Outcome for OS

OS is defined as the time from diagnosis or treatment initiation to death from any cause. The survival data includes right-censoring for patients alive at last follow-up, requiring models to account for incomplete event times. Survival data offers three essential pieces of information for each instance or patient: observed features, time elapsed since features were first collected, and a label indicating whether the event (e.g., death) has occurred. In our approach, we consider survival time as discrete, with a finite time horizon. The time set, denoted as T, is defined as $T = \{0,...,T_{max}\}$, where T_{max} represents the predetermined maximum time horizon. Given that the event of interest may not always be observed due to patients being lost to follow-up, survival data often involve truncation. Truncation happens when the observation of an individual ends before the event of interest occurs. Patients may cease participating in the study or being monitored before the event of interest takes place. Addressing this challenge is a crucial aspect of our analysis. We define *truncation* as the event 0 and the set of possible events, including truncation, as $K = \{0, 1\}$, where 1 represents the event of interest. Each data point or instance is therefore a triple ($\mathbf{x}$, s, k) where $\mathbf{x} \in X$ is a D-dimensional vector of features, $s \in T$ is the time at which the unique event or

truncation occurred, and $k \in K$ is the event or truncation that occurred at time s. The dataset $D = \{(\mathbf{x}^{(i)}, s^{(i)}, k^{(i)})\}_{i=1}^{N}$ describes a finite set of observed instances or patients in our analysis. For each tuple $(\mathbf{x}^{(i)}, s^{(i)}, k^{(i)})$ with $k^{(i)} \neq 0$ our focus lies in determining the actual probability $P(s = s^{(i)}, k = k^{(i)} | \mathbf{x} = \mathbf{x}^{(i)})$. This probability models the likelihood that a patient with features $\mathbf{x}^{(i)}$ will encounter the event $k^{(i)}$ at the specific time $s^{(i)}$. Given the inherent limitation that the true probability cannot be precisely ascertained from any finite dataset, our objective is to derive estimates $\hat{P}$ that serve as approximations to these true probabilities.

Risk Assessment Network The body of literature addressing survival analysis often approaches the event of interest as the first hitting time of an underlying stochastic process. In a medical context, survival analysis pertains to the duration a patient survives. A significant challenge in survival analysis involves understanding the relationship between the distribution of hitting times and the covariates, which represent individual features. Previous research in this field often assumes a specific form for the underlying stochastic process, utilizing available data to learn the relationship between covariates and the parameters of the model, subsequently deducing the connection between covariates and the distribution of first hitting times, also known as the risk of the event (e.g., the risk of death).

We take a markedly different approach to survival analysis by leveraging a deep neural network named DeepHit [86] (Figure 3.3), which allows us to model the complex, nonlinear relationships between high-dimensional patient representations and survival outcomes without imposing strong parametric assumptions on the underlying stochastic process. This flexibility enables the network to more accurately capture patient-specific risk patterns and to directly integrate rich, semi-structured embeddings derived from clinical narratives.

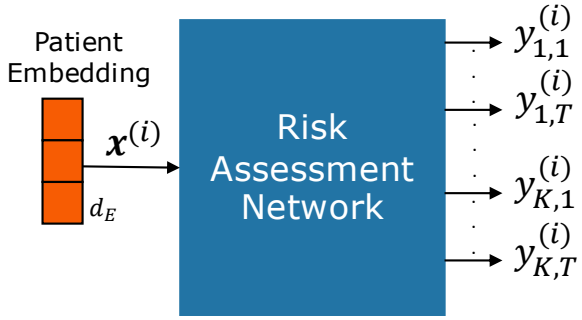

Figure 3.3: OS prediction: The outcome of HEAL x(i) is the input of the risk assessment network DeepHit.

DeepHit learns the distribution of first hitting times directly, without making assumptions

about the form of the underlying stochastic process. The objective is to instruct the risk assessment network to acquire the knowledge of $\hat{P}$, the estimate for the joint distribution of the first hitting time and competing events. DeepHit consists of a shared sub-network (SN) and multiple cause-specific sub-networks (CSNs), contingent upon the number of events k. To guarantee the learning of the joint distribution of k competing events, as opposed to the marginal distributions of individual events, DeepHit employs a single softmax layer as its output layer. Furthermore, the model incorporates a residual connection linking the input covariates to the input of each cause-specific sub-network, contributing to the overall robustness and effectiveness of the learning process. In our specific context, the sole event under consideration is the patient's death, denoted as $k = 1$. Consequently, we have just one cause-specific subnetwork. The shared sub-network and the cause-specific sub-network are composed of L_S and L_C fully-connected layers, respectively. The shared sub-network takes clinical covariates $\mathbf{x}^{(i)}$ as inputs and generates an output vector $f_s(\mathbf{x}^{(i)})$ capturing the latent representation of covariates. On the other hand, the cause-specific sub-network takes pairs $z = (f_s(\mathbf{x}^{(i)}), \mathbf{x}^{(i)})$ as inputs and produces an output vector $f_c(z)$ representing the probability of the first hitting time. Notably, these sub-networks incorporate both the output of the shared network and the original covariates as inputs. This design choice allows the sub-networks to access the learned common representation $f_s(\mathbf{x}^{(i)})$ while retaining the ability to learn distinct aspects of the representation. The softmax layer generates a probability distribution denoted as $\mathbf{y} = [y_{1,1}, ..., y_{1,T_{max}}]$, where $y_{1,s}$ represents the estimated probability $\hat{P}(s^{(i)}, k^{(i)}|\mathbf{x}^{(i)})$ indicating the likelihood of the patient experiencing event $k^{(i)} = 1$ at the time $s^{(i)}$. This architectural framework encourages the network to grasp potentially non-linear and even non-proportional relationships between covariates and associated risks. To assess the risk of event occurrence, the cause-specific cumulative incidence function (CIF), expressed as $F_{k^{(i)}}(s^{(i)}|\mathbf{x}^{(i)})$, is employed. This function quantifies the probability of the event $k^{(i)} = 1$ occurring on or before time $t^{(i)}$, given the covariates $\mathbf{x}^{(i)}$. Formally, the CIF for event $k^{(i)} = 1$ is expressed as:

$$F_{k^{(i)}}(t^{(i)}|\mathbf{x}^{(i)}) = \sum_{s^{(i)}=0}^{t^{(i)}} P(s = s^{(i)}, k = k^{(i)}|\mathbf{x} = \mathbf{x}^{(i)}). \tag{3.3}$$

However, since the true CIF $F_{k^{(i)}}(s^{(i)}|\mathbf{x}^{(i)})$ is not known, we utilize the estimated CIF:

$$\hat{F}_{k^{(i)}}(t^{(i)}|\mathbf{x}^{(i)}) = \sum_{m=0}^{s^{(i)}} y_{1,m}. \tag{3.4}$$

Loss Function In our implementation of DeepHit we define a loss function $\mathcal{L}$, that has been specifically crafted to effectively handle censoring data. It is expressed as the formula $\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3$ where α, β, γ weight the three terms $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ now described. The term $\mathcal{L}_1$ embodies the log-likelihood of the joint distribution concerning the first hitting time and the unique event, i.e. death ($k^{(i)} = 1$). Notably, this formulation has been adapted to accommodate the presence of censored data. For patients who have not experienced censoring, $\mathcal{L}_1$ encapsulates both the occurrence of the event and the corresponding time at which it occurred. On the other hand, for patients who have been subject to censoring, $\mathcal{L}_1$ effectively captures the time at which the patient becomes censored, indicating that they were alive up to that specific point in time and providing valuable information regarding their status at that juncture. This adjustment ensures that censoring is appropriately accounted for, offering a more accurate representation of patient outcomes. Formally

$$\begin{aligned} \mathcal{L}_1 = -\sum_{i=1}^{N} [\mathbb{1}(k^{(i)} \neq 0) \cdot log(y^{(i)}_{k^{(i)},s^{(i)}}) \\ +\mathbb{1}(k^{(i)} = 0) \cdot log(1 - \sum_{k=1}^{K} \hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}))], \end{aligned} \tag{3.5}$$

where $\mathbb{1}()$ is an indicator function and N is the number of patients in the dataset. The first term captures the information contributed by patients who have not undergone censoring. The second term addresses censoring bias by leveraging the understanding that these patients are confirmed to be alive at the time of censoring. This acknowledgment enables the model to anticipate that the first hitting event will occur after the specified censoring time.

$\mathcal{L}_2$ integrates a blend of cause-specific ranking loss functions, and it uses the estimated CIFs computed at various times, corresponding to the instances when events actually occur. This approach is employed to fine-tune the network for each cause-specific estimated CIF. Since this study focuses on a single event, there is only one cause-specific estimated CIF. Our methodology employes a ranking loss function that incorporates the concept of concordance: a patient experiencing an event at time s should exhibit a higher risk at that specific time s than a patient who has survived beyond s. This ensures that the model not only predicts the occurrence of the event, but also correctly orders the risks of death over time. Formally

$$\mathcal{L}_2 = \sum_{k=1}^{K} \theta_k \cdot \sum_{\substack{i=1 \\ i \neq j}}^{N} A_{k,i,j} \cdot \eta(\hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}), \hat{F}_k(s^{(i)}|\mathbf{x}^{(j)})), \tag{3.6}$$

where the coefficients θ_k are chosen to trade off ranking losses of the $k-th$ competing event, $\eta(a,b)$ is a convex loss function defined as $\eta(a,b) = \exp\left(-\frac{(a-b)}{\sigma}\right)$ with σ set equal to 0.1 and

$A_{k,i,j}$ is defined as follows:

$$A_{k,i,j} = \mathbb{1}(k^{(i)} = k, s^{(i)} < s^{(j)}), \tag{3.7}$$

and represents pairs (i, j) acceptable for event k. Since this study focuses on a single event, only one coefficient, i.e. θ_1, is included, and its value is fixed at 1. The inclusion of $\mathcal{L}_2$ in the overall loss function penalizes the misordering of pairs concerning each event. Consequently, minimizing the total loss serves to incentivize the correct ordering of pairs for each event.

$\mathcal{L}_3$ is a calibration loss: it focuses on how well predicted probabilities align with observed outcomes, ensuring that the model's predicted risk accurately reflects the true event occurrence. It is defined as follows:

$$\mathcal{L}_3 = \sum_{k=1}^{K} \frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}) - I_i), \tag{3.8}$$

where I_i represents the indicator of the event, specifically, the death of the patient. When I_i equals 1, it signifies the occurrence of the patient's death. Conversely, when I_i equals 0, it indicates truncation.

Binary Treatment Response Endpoint for pCR

Pathologic complete response (pCR) was defined as the absence of residual invasive cancer in both the primary tumor and the regional lymph nodes following neoadjuvant therapy, as confirmed by histopathological examination after surgery. This endpoint is widely recognized as a surrogate marker of favorable long-term outcomes in several cancer types.

The patient representations generated by HEAL were used as input to a prediction module designed to address the binary nature of the pCR task. Given the limited number of available patients and the consequent risk of overfitting associated with high-capacity models, we explored both deep learning and classical machine learning approaches.

First, a neural network classifier implemented as a three-layer multi-layer perceptron (MLP) was employed to model non-linear relationships between the learned patient representations and treatment response. The MLP outputs class probabilities corresponding to the two possible outcomes (pCR vs. no-pCR), as illustrated in Figure 3.4. To mitigate class imbalance, Focal Loss was adopted during training, and optimization was performed using standard stochastic gradient-based methods.

In addition to deep learning, we evaluated several classical machine learning classifiers operating on the same HEAL-derived patient representations. These models, characterized by lower capacity and stronger inductive biases, are often better suited for small-sample clinical settings. In particular, linear and non-linear machine learning methods were included

to assess whether simpler decision boundaries could yield improved generalization compared to deep architectures in this low-data regime.

This dual evaluation strategy allows us to disentangle the contribution of representation learning from that of the final classifier. While the attention-based HEAL framework is responsible for extracting clinically meaningful and compact patient-level embeddings from unstructured text, the downstream prediction can be flexibly adapted to the task and data characteristics.

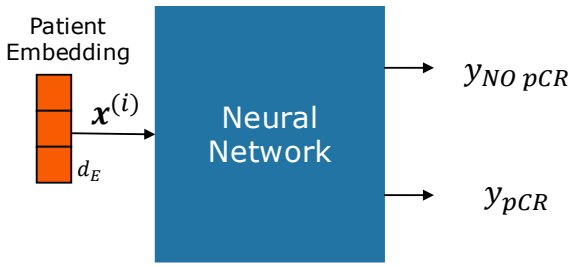

Figure 3.4: pCR prediction pipeline: patient representations produced by HEAL serve as input to a neural network classifier.

3.3 OS Experiments and Results

3.3.1 Experimental Setup

For OS prediction, within each cross-validation iteration, a stratified inner 10-fold cross-validation was conducted to fine-tune the network hyperparameters reported in Table 3.2. This optimization process involved a random search with 100 iterations over the hyperparameter space depicted in Table 3.2. The entire process was repeated five times for each setting (HEAL, clinical features, ablation tests) to address fluctuations in the results, ultimately providing a more reliable and precise perspective on the performance of the risk assessment network. All experiments were implemented in PyTorch and run on a NVIDIA A100 GPU with 40 GB of VRAM.

Baselines and Ablation Strategy

This section outlines the baseline comparisons and ablation studies designed to rigorously evaluate the effectiveness and robustness of our proposed approach.

Table 3.2: Hyperparameters search space of the risk assessment network.

Hyperparameter		**Search space**
Batch size		[8, 16]
# hidden layers for both SN and CSNs		[1, 2, 3, 5]
# neurons per hidden layer		[20, 50, 100, 200]
Dropout rate		[0.2, 0.3, 0.4]
Activation function		[ReLU, SELU]
Loss function $\mathcal{L}$	α	[0.1, 0.5, 1.0, 3.0]
	β	[0.1, 0.5, 1.0, 3.0]
	γ	[0.1, 0.5, 1.0, 3.0]

As a first step, we benchmarked our method against a set of clinically relevant features manually extracted by domain experts, as reported in Table 3.3. These features were defined under the guidance of two specialists, an oncologist and a radiation oncologist, and have been employed in prior research [20]. In this baseline configuration, they served as the exclusive input to the risk assessment network, providing a clinically meaningful reference for comparison.

Table 3.3: Patients' features.

Feature	**Description**	**Values**
Gender	The gender of the patient	[M, F]
Overall Stage	The cancer stage, indicating the extent of the disease progression	[II, III, IV]
cT	The clinical tumor classification, providing the tumor size information	[1, 2, 3, 4]
cN	The lymph node classifications, providing information about the lymph node involvement	[0, 1, 2, 3]
cM	The metastasis classification, providing information about metastatic status	[0,1]
Histology	The specific diagnosis related to the cancer type	[Adenocarcinoma, Squamous, Other, Unknown]
CTV	This feature pertains to the volume of tissue that is outlined for therapeutic irradiation	[1.8–568.61] cm^3
Age	The age of the patient	[29–92]

Beyond this baseline, we carried out a series of ablation studies to disentangle and quantify the contribution of each component of our method. Specifically, we investigated the following configurations:

- **No NER**: this approach involves combining at first attention layer (token level attention in Figure 3.2) the embedding of all tokens within a sentence, not only of those

belonging to an entity type, ignoring the NER output. It serves to assess the importance of incorporating NER classification output to understand the contribution of NER classification in order to extract discriminative features for OS prediction.

- **Binary NER**: in binary NER, a token within a sentence is classified as belonging to an entity or not, without additional subcategorization into specific types defined in Table 2.2. It serves to evaluate the relevance of multiclass NER as opposed to binary NER in order to extract discriminative features for OS prediction.

- **No HEAL**: this approach involves substituting HEAL with a simple average among all the entity token embeddings that are outputted by the NER system, without any distinction in sentences. It allows to assess the importance of weighting information in clinical reports.

- **Only Sentence Attention (SA):** this approach substitutes only the first attention layer (token level attention in Figure 3.2) with a simple average among all entity token embedding of a sentence. The sentence level attention is maintained, allowing to understand the importance of weighting sentences within patient clinical reports.

- **Only Token Attention (TA):** this approach substitutes the second attention layer (sentence level attention in Figure 3.2) with a simple average among all sentence token embedding in the patient clinical reports. The token level attention is maintained, allowing to understand the importance of weighting tokens within sentences.

- **NER degradation**: this method utilizes a degraded NER system with an F1 score of 65%, achieved by introducing noise through the random removal of correctly identified entities. This approach allows for the evaluation of how a reduction in entity quality impacts overall model performance.

3.3.2 Results and Discussion

We use the time-dependent concordance index (C^{td}-index) as metric of performance, which ranges from 0 to 1. It is important to highlight that the conventional concordance index (C-index) [51] is a widely utilized discriminative metric. The C-index operates under the assumption that patients with longer lifespans should be associated with a lower risk compared to those with shorter lifespans. However, the ordinary C-index is calculated solely at the initial observation time, lacking the capacity to capture potential variations in risk over time. In contrast, the time-dependent concordance index considers the temporal aspect, offering a

more comprehensive understanding of how risk evolves over the course of observation. The C^{td}-index for event k is defined as:

$$\begin{aligned} C^{td} &= P(\hat{F}_k(s^{(i)}|x^{(i)}) > \hat{F}_k(s^{(i)}|x^{(j)})|s^{(i)} < s^{(j)}) \\ &\approx \frac{\sum_{i \neq j} A_{k,i,j} \cdot \mathbb{1}(\hat{F}_k(s^{(i)}|x^{(i)}) > \hat{F}_k(s^{(i)}|x^{(j)}))}{\sum_{i \neq j} A_{k,i,j}}. \end{aligned} \tag{3.9}$$

Thus, the C^{td}-index for event k is computed by comparing pairs of observations. In each pair, one patient has experienced event k at a specific time, whilst the other has neither encountered the event nor been truncated to that time. The significance of this discriminative index lies in its independence from a single fixed time. This characteristic renders it well-suited for situations where the impact of covariates on survival undergoes variations over time. In other words, this index is particularly valuable when risks exhibit non-proportional behavior over the course of observation.

Table 3.4 summarizes the results averaged over 5 runs, presenting the performance metrics in terms of the C^{td}-index for the compared modality. For the HEAL modality, the model achieved an average C^{td}-index of 0.639 with a low standard deviation of 0.014, which is lower than the standard deviations of other methods, indicating higher consistency across the runs. Conversely, the model's performance decreased in No HEAL modality, yielding an average C^{td}-index of 0.558. This observation highlights the importance of appropriately weighing information within clinical reports for optimal predictive outcomes. The ablation test with only the Sentence Attention (SA) mechanism led to improved performance, with an average C^{td}-index of 0.624. Although slightly lower than HEAL, this enhancement suggests that strategically weighting sentences inside clinical reports had a positive impact to overall performance. In the absence of weight sharing inside HEAL, the model exhibited an average C^{td}-index of 0.615, which is slightly lower than HEAL by 0.009. This suggests that the network performs better with a reduced number of parameters, possibly due to training with a limited number of samples. Both the binary NER (Bi-NER) and the no NER modalities resulted in significatively lower performances, with an average C^{td}-index of 0.546. This indicates the fundamental role of NER label information in training an effective predictive model, emphasizing that tokens not associated with entities are of low informational value. Clinical features exhibited a slightly lower performance with an average C^{td}-index of 0.590. This suggests that the proposed automated process outperforms manually extracted features by humans. In summary, these results offer insights into the relative effectiveness of different modalities and model configurations in predicting risk, with the HEAL modality emerging as the most consistent and effective among the tested approaches.

In order to further validate the difference between the proposed approach and the com-

Table 3.4: Results of C^{td}-index for individual modalities obtained from 5 iterations.

Approach	C^{td}**-index (mean ± std)**
Clinical features	0.590 ± 0.019
No NER	0.546 ± 0.029
Bi-NER	0.499 ± 0.023
No HEAL	0.558 ± 0.023
SA	0.624 ± 0.027
TA	0.570 ± 0.037
No WS	0.615 ± 0.033
HEAL	$\mathbf{0.639 \pm 0.014}$

pared methods, we performed the Student's t-test in a pairwise fashion, considering HEAL angaist each competitor. The results are summarized in Table 3.5. A significance threshold $\hat{\alpha}$ of 0.05 was established for the conducted tests. The results of t-tests again highlight the pivotal role of attention mechanisms:for all competitors lacking attention mechanisms, the p-values consistently fell below the established threshold of 0.05, indicating statistical significance. The significance of the p-value persists even for the TA competitor, where an attention mechanism is present. However, this mechanism aggregates all words in patient clinical reports, disregarding sentence splitting and thus the hierarchical structure of our methodology. Noteworthy are the elevated p-values associated with the other scenarios featuring at least one attention mechanism, specifically in the exclusive presence of the Sentence Attention (p-value = 0.290) and the absence of weight sharing (p-value = 0.174). Even if the two approaches are similar to HEAL, the C^{td}-index mean and standard deviation values reveal a consistent trend towards superior and more robust outcomes with HEAL. Attributing our model's superiority, we highlight two key factors: the inclusion of weighting tokens within sentences before weighting the sentences themselves, and the use of weight sharing. The first significantly enhances the comprehension of clinical information, particularly when dealing with larger data volumes compared to the Sentence Attention (SA) approach. The second reduces the number of parameters compared to the No Weight Sharing (No WS) approach, contributing to the model's efficiency and effectiveness.

3.3.3 Interpretability and Clinical Insights

HEAL provides attention maps at both token and sentence levels, enabling inspection of which clinical concepts influenced predictions. An example of attention map is shown in

Table 3.5: Statistical analysis of performance differences between HEAL and the other modalities. Statistically significantly differences (p-value<0.05) are highlighted in bold.

Approach	**Compared to**	**ΔC^{td}-index (mean)**	**p-value**
HEAL	Clinical features	0.049	**0.001**
	No NER	0.093	**<0.001**
	Bi-NER	0.140	**<0.001**
	No HEAL	0.081	**<0.001**
	SA	0.015	0.290
	TA	0.069	**0.004**
	No WS	0.024	0.174

Figure 3.5, which presents sentences extracted from the clinical reports of a patient who has been assigned an 86% risk score of experiencing " death " within 29 months.

To evaluate the robustness and reliability of the model in generating attentional maps we implemented a sanity check that involved comparing attentional maps obtained from a faulty system angaist those from the accurate system. The attentional maps from the faulty system are generated by training the model using randomly permuted OS labels, following a data randomization test [3]. The primary objective of the sanity check is to discern whether the model can distinguish between the true signals influencing its decisions and mere noise or random associations. Cosine similarity analysis between attentional maps of the two systems yielded a mean score of 0.578 with a standard deviation of 0.138, indicating a relatively substantial difference between attentional maps from the faulty and accurate systems.

Furthermore, to qualitatively assess the attentional maps generated by our model during the decision-making process, we measures the agreement with domain experts on the importance given by the model to the sentences within patients' clinical reports for predicting clinical outcomes. To this end we set up a questionnaire consisting of four questions, each corresponding to an individual patient's clinical report under examination. For each question, participants assess the level of agreement with the significance assigned by the model to the sentences within the report. In particular, each sentence in the report was highlighted with a specific color denoting its importance. We employed a three-level highlighting system: orange for the most important sentences (those receiving the highest attentional weights from the model), blue for the least important sentences (those receiving the lowest attentional weights from the model) and green for sentences falling in between. For each question, participants have the option to choose from five different response levels: completely agree, agree, neutral, disagree, and completely disagree. These responses were systematically encoded on a numer-

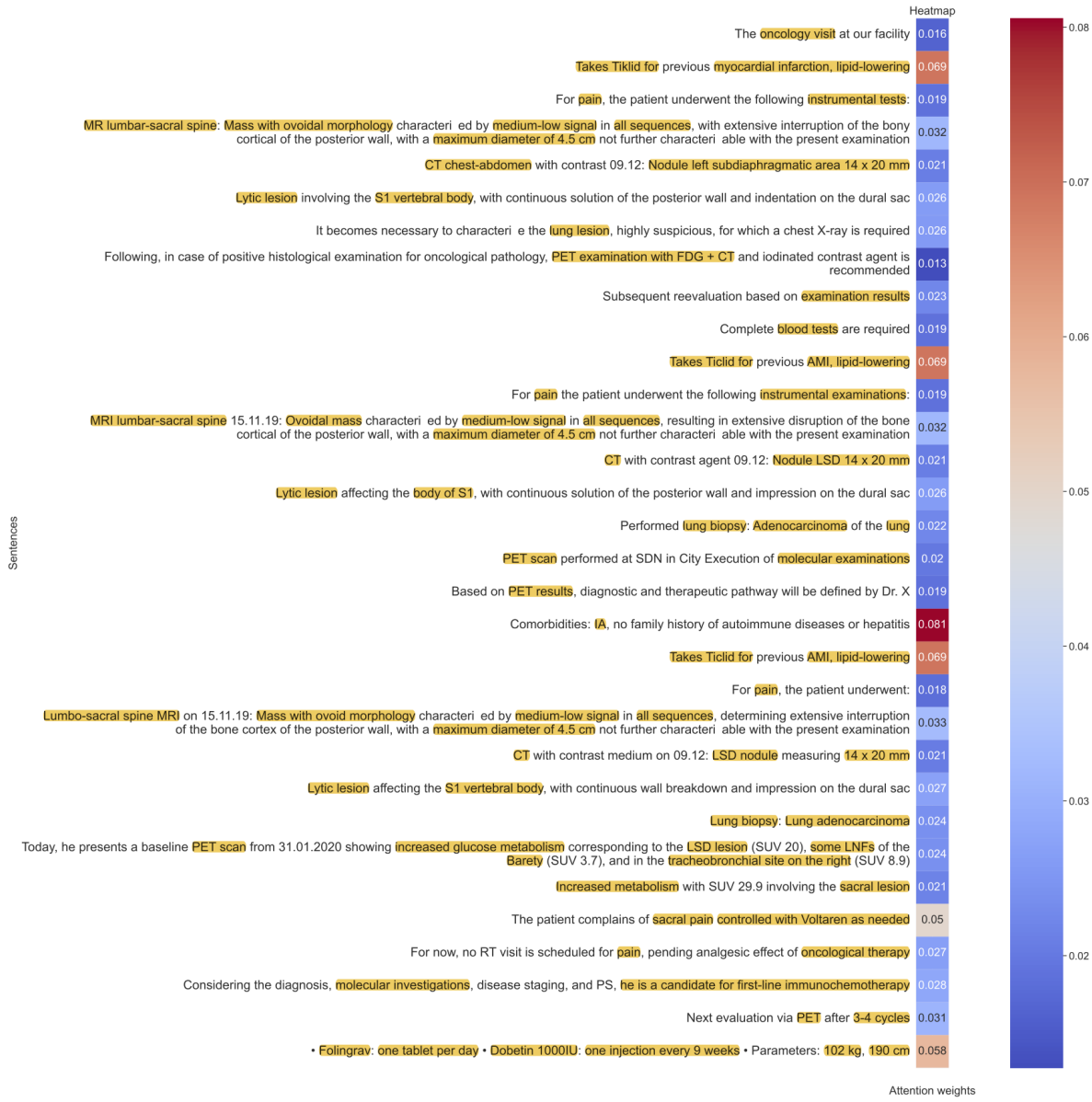

Figure 3.5: OS prediction: The outcome of HEAL x(i) is the input of the risk assessment network DeepHit.

ical scale using the Likert scale, with values ranging from 0 for "completely disagree" to 4 for "completely agree". The questionnaire was proposed to four domain-experts. We obtained an overall agreement of 67.2%, which suggests a substantial level of consensus among the respondents. This indicates a noteworthy degree of alignment in perceptions regarding the model's attention to critical information within clinical reports. However, it is important to further explore the remaining 32.8%, where participants have divergent views. Specifically, these divergent perspectives manifest a clinicians' tendency to attribute higher significance

to particular sections in the descriptions of CT and MRI exams, as well as to the diagnosis itself, even though the model does not categorize these details as the least crucial in the reports. Moreover, since the diagnosis is consistent for almost all patients in the cohort, the model may not emphasize this feature significantly when distinguishing between patients for prognostic purposes.

3.4 pCR Experiments and Results

3.4.1 Experimental Setup

The experimental evaluation was designed to assess the impact of different architectural choices within the proposed deep learning framework, as well as to compare its performance against traditional machine learning baselines, particularly in light of the limited number of available patients.

Deep Learning configurations For the deep learning experiments, multiple network configurations were explored by varying both the attention mechanism and the classification layers. The goal was to analyze how different levels of model expressiveness and inductive bias affect performance under data-constrained conditions.

Specifically, the attention component was instantiated using three alternative designs: convolutional layers, fully connected (FC) layers, and multi-layer perceptrons (MLPs).

Similarly, the classification module following the attention mechanism was implemented using either a single FC layer or a deeper MLP. This choice allows controlling the trade-off between model capacity and overfitting risk, which is particularly relevant given the relatively small cohort size.

All deep learning models were trained using the Focal Loss to mitigate class imbalance, the Adam optimizer, a batch size of 2, and a learning rate of 0.001. These hyperparameters were kept fixed across configurations to ensure a fair comparison between architectural variants.

Machine Learning baselines In addition to deep learning models, traditional machine learning (ML) approaches were included as comparative baselines. Given the limited number of patients, classical ML methods often provide strong performance due to their lower variance and reduced susceptibility to overfitting. These models operate on the same extracted feature representations and serve as a reference to assess whether the increased complexity of deep architectures yields consistent performance gains in low-data regimes.

Implementation details All experiments were implemented in PyTorch and executed on an NVIDIA A100 GPU with 40 GB of VRAM.

3.4.2 Results

Evaluation Criteria

The performance of the model was assessed using several metrics: Accuracy, AUC, F1 Score, and MCC. These metrics were chosen to provide a comprehensive evaluation of the model's performance, particularly given the strong class imbalance present in the dataset (77% positive vs. 33% negative).

Accuracy is the ratio of correctly predicted instances, which include True Positives (TP) and True Negatives (TN) representing the correctly classified positive and negative classes, respectively, to the total number of instances, including False Positives (FP) and False Negatives (FN), representing the misclassified positive and negative classes, respectively. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

The AUC is a measure of the ability of the model to distinguish between positive and negative classes. It represents the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR, also known as Recall, measures the proportion of actual positives that are correctly identified by the model, while the FPR measures the proportion of actual negatives that are incorrectly identified as positives by the model.

The F1 score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns. It is particularly useful when the class distribution is imbalanced. The Precision, Recall and F1 score are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3.11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.12}$$

The MCC is a measure of the quality of binary classifications, taking into account true and false positives and negatives, and is generally regarded as a balanced measure even for imbalanced datasets. MCC is given by:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.13}$$

An MCC of $+1$ indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates total disagreement between prediction and observation.

Results and Discussion

The results are reported in Table 3.6 and Table 3.7. As we can see from Table 3.6, using an FC as a classification layer lead to the best accuracy of 64.2% by using convolutional layers within the attentional mechanisms. The real effectiveness compared with using convolutional, FC or MLP layers can be deduced from the MCC index, on which, on the other hand, a higher difference is observed. The use of FC layer achieves the best value, which is 10.4%. In contrast to Table 3.6, the use of an MLP as a classification layer brings all the best performance on the use of convolutional layers in the dual attention mechanism, with an accuracy of 73.1% and a MCC of 22.7%, as shown in Table 3.7. In order to improve the performance of the latter results, a reduction in embedding size and the use of an optimized AdamW with weight decay was tested. The results are shown in Table 3.8.

In Table 3.8 we note the improvement with an embedding size of 768 achieving an MCC of 23.9%, bringing a +1.2% improvement over the parameters in Table 3.7. Given the small number of patients, the deep learning approach did not achieve satisfactory performance. For this reason a classical classification approach using machine learning was tested. Specifically, the algorithms tested are: Decision Trees [156], SVM [31], Random Forest [55] and XGBoost [25].

As shown in Table 3.9, the machine learning-based approach allowed a large jump of about +13.2% on the MCC index and +1.5% on accuracy compared with the best result from deep learning techniques. Given the promising results obtained, in order to concretely validate the effectiveness of encoding the dual attention mechanism, in addition to the features extracted from the EHRs, clinical features extracted from the 67 patients were added, and the results were compared with the experiments performed without the clinical features.

By comparing Table 3.9, 3.10 and 3.11 emerges the actual informational contribution made by each feature and its combinations. The EHR features alone produced by the double attention mechanism bring a great improvement over the deep learning models. The clinical features alone come close to this result with the Random Forest [55], achieving an accuracy of 80.6% and an MCC of 33.3%, a +1.5% and −3.8%, respectively, compared with the EHR features. The combination of these clinical features and those of EHRs led to the results shown in Table 3.11. These performances see the XGBoost [25] algorithm as the best on all

metrics considered. Moreover, these performances are the same as those seen in Table 3.9 considering only the features extracted from the EHRs, which is not the case for the other algorithms, that see a slight drop in performance. The best MCC index is, again, 37.1%.

Table 3.6: Deep Learning with a Fully Connected classification layer.

Metric	Convolution	FC	MLP
Accuracy	**0.642**	0.612	0.582
AUC	0.485	**0.560**	0.517
F1	0.200	**0.350**	0.300
MCC	−0.030	**0.104**	0.029

Table 3.7: Deep Learning with a Multi-layer Perceptron classification layer.

Metric	Convolution	FC	MLP
Accuracy	**0.731**	0.716	0.641
AUC	**0.613**	0.556	0.556
F1	**0.400**	0.300	0.333
MCC	**0.227**	0.122	0.100

Table 3.8: Deep Learning results using the Multi-layer Perceptron as classification layer with AdamW and weight decay as the embedding size varies.

Metric	128-d	256-d	768-d
Accuracy	0.746	**0.776**	0.716
AUC	0.576	0.571	**0.628**
F1	0.320	0.286	**0.424**
MCC	0.177	0.208	**0.239**

Table 3.9: Machine Learning results using the features of the EHRs.

Metric	Decision Trees	SVM	Random Forest	XGBoost
Accuracy	0.582	0.582	**0.791**	0.776
AUC	0.446	0.494	0.604	**0.689**
F1	0.176	0.263	0.363	**0.516**
MCC	−0.099	−0.011	0.284	**0.371**

Table 3.10: Machine Learning results using the Clinical features.

Metric	Decision Trees	SVM	Random Forest	XGBoost
Accuracy	0.731	0.522	**0.806**	0.701
AUC	0.613	0.455	0.614	**0.618**
F1	0.400	0.238	0.381	**0.412**
MCC	0.227	−0.076	**0.333**	0.218

Table 3.11: Machine Learning results using the features of the EHRs and the Clinical features.

Metric	Decision Trees	SVM	Random Forest	XGBoost
Accuracy	0.657	0.582	0.761	**0.776**
AUC	0.542	0.494	0.562	**0.689**
F1	0.303	0.263	0.273	**0.516**
MCC	0.080	−0.011	0.168	**0.371**

3.5 Cross-task Discussion and Chapter Summary

This chapter evaluated the proposed HEAL framework across two clinically distinct prediction tasks: OS, formulated as a time-to-event problem, and pCR, modeled as a binary classification task. Although both tasks rely on the same underlying representation learning

strategy based on NER-driven hierarchical attention, their experimental outcomes highlight important differences in task characteristics, data availability, and modeling requirements.

For OS prediction, HEAL demonstrated strong and consistent performance improvements over all baseline configurations, including manually engineered clinical features and multiple ablated variants. The hierarchical attention mechanism proved particularly effective in this setting, as evidenced by both quantitative gains in the time-dependent concordance index and qualitative interpretability analyses. These results suggest that survival prediction benefits substantially from the ability of HEAL to selectively weight clinically relevant entities and sentences across longitudinal narratives, capturing nuanced temporal and semantic patterns that are difficult to encode through structured variables alone. Moreover, the reduced variance observed across runs indicates that the architecture generalizes well despite the moderate sample size.

In contrast, the pCR task presented a markedly different scenario. The limited number of patients and the binary nature of the endpoint constrained the effectiveness of deep learning models, even when architectural refinements and regularization strategies were applied. While the dual attention mechanism still contributed meaningful representations, classical machine learning methods consistently outperformed deep architectures in this low-data regime. This behavior aligns with well-known bias–variance trade-offs: simpler models with lower capacity are often better suited when data scarcity limits the reliable estimation of complex decision boundaries.

Importantly, the pCR experiments also demonstrated that features derived from HEAL remain highly informative, as machine learning models operating on these representations achieved the best overall performance. This finding confirms that the proposed attention-based encoding captures clinically relevant signals, even when the final prediction is delegated to non-deep classifiers. In this sense, HEAL acts as a powerful feature extraction and representation learning module, adaptable to different downstream modeling paradigms depending on task complexity and data availability.

Taken together, these results highlight the flexibility of the HEAL framework across heterogeneous clinical prediction tasks. For time-to-event outcomes with richer supervision and longer follow-up, end-to-end deep survival modeling fully exploits the benefits of hierarchical attention. Conversely, for small-sample binary endpoints, combining HEAL-derived representations with classical machine learning offers a more effective and robust solution. This dual behavior underscores the general applicability of HEAL as a modular and task-aware approach for learning from unstructured clinical text.

Chapter 4

Slice-Level Attention for Survival Prediction from Chest CT Imaging

Chest CT is a cornerstone of NSCLC management, routinely employed for diagnosis, staging, and response evaluation. Beyond its diagnostic role, CT imaging holds rich prognostic information, as tumor size, morphology, and intra-tumoral heterogeneity have all been associated with survival outcomes. Traditionally, this prognostic value has been extracted through manual interpretation or radiomics, the latter relying on handcrafted features quantifying shape, intensity, and texture. However, these approaches face intrinsic limitations: manual assessments are prone to inter-observer variability, whilst radiomics requires carefully engineered features that may fail to capture the full complexity of tumor biology.

Deep learning has emerged as a transformative alternative, enabling the automatic extraction of high-dimensional features directly from imaging data [101]. CNNs have achieved remarkable success in medical image analysis tasks such as detection, segmentation, and classification. Yet, their application to survival prediction remains challenging. A primary obstacle is the volumetric nature of CT scans: although 3D CNNs can process entire volumes, they are computationally expensive and prone to overfitting in data-limited scenarios. Conversely, 2D CNNs process slices independently, but lack mechanisms to integrate slice-level information into coherent volume-level prognostic representations.

To address this limitation, a slice-level attention mechanism is introduced to explicitly model the contribution of individual slices within a CT volume. Rather than treating all slices as equally informative, the model learns to emphasize those that carry stronger prognostic signals, while down-weighting less relevant or redundant slices. This design enables the use of efficient 2D convolutional networks while still capturing volumetric context through attention-based aggregation. By selectively weighting slices according to their relevance for survival prediction, the resulting patient-level representation becomes more compact, with

an implicit focus on tumor-bearing regions.

This chapter presents a slice-level attention framework for survival prediction in NSCLC, integrating an EfficientNetB0 encoder for slice-level feature extraction with a soft attention mechanism to aggregate slice representations into a comprehensive volume-level embedding. This representation is subsequently processed by a survival analysis network to predict patient-specific risk across a 24-month horizon. We describe the preprocessing pipeline, model design, experimental setup, and results on two datasets, LUNG1 and CLARO, highlighting the advantages of slice-level attention over conventional 3D CNNs and slice-agnostic strategies.

4.1 Motivation and Clinical Background

CT is the most widely adopted imaging modality in the clinical management of NSCLC. It provides high-resolution anatomical detail of lung parenchyma, mediastinum, and tumor morphology, making it the gold standard for initial diagnosis, staging, treatment planning, and follow-up. In particular, CT imaging allows radiologists to assess tumor size, location, nodal involvement, and metastatic spread, which are critical prognostic factors.

Beyond anatomical assessment, CT has been shown to correlate with patient outcomes, as quantitative imaging features, such as tumor volume, shape irregularity, and intra-tumoral heterogeneity, are linked to survival. However, prognostic inference from CT is traditionally dependent on manual interpretation and handcrafted radiomics [22], both of which are subject to variability, inter-observer bias, and limitations in capturing subtle prognostic cues. Deep learning–based approaches promise to overcome these challenges by automatically extracting hierarchical representations from raw CT data [101]. Recent studies [14, 171, 116, 150, 48] have explored the use of deep learning for OS prognosis in NSCLC patients using the public NSCLC-Radiomics (LUNG1) dataset from the MAASTRO clinic [4], which, to the best of our knowledge, represents the most comprehensive and extensive resource for OS prediction in NSCLC cases. In Braghetto et al. [14], radiomics and deep-learning approaches were compared on the LUNG1 dataset for the classification of 2-year OS. For radiomics, the study considered the best combination between two feature selectors (ANOVA, Cluster reducer) and six classifiers (SVM, BAG, XGB, Neural Network, KNN, RF). The optimal combination achieved an average AUC of 0.67 across five random test splits. The deep learning approach, consisting of a 2D convolutional neural network (CNN), achieved a slightly lower average AUC of 0.64 across the same splits. In Zheng et al. [171], a hybrid model that integrated both image and clinical features was implemented using a 3D CNN for the classification of 2-year OS in stage I-IIIA non-small cell lung cancer pa-

tients. Image features were learned from cubic patches containing lung tumors extracted from pre-treatment CT scans. Relevant clinical variables were identified through analyses, revealing that age and clinical stage are the most significant prognostic factors for 2-year OS. Using these two clinical variables in combination with image features from pretreatment CT scans, the hybrid model, after training on the University Medical Center Groningen (UMCG) dataset, achieved a median AUC of 0.64 on the LUNG1 test set, which contained 228 patients (n=228) with stage I-IIIA lung tumors treated with radiation or concurrent chemoradiation. In [116], a foundation model for cancer 2-year OS classification was developed by training a convolutional encoder through self-supervised learning using a comprehensive dataset of 11,467 radiographic lesions. The foundation model was then fine-tuned using data from the HarvardRT cohort (n=291) to classify 2-year OS after treatment. Subsequently, the model was evaluated on the LUNG1 cohort, achieving an AUC of 0.638. In [150] Torres et al. developed a fully automated imaging-based prognostication technique (IPRO) using a 3D CNN to predict 1-year, 2-year, and 5-year mortality from pretreatment CTs of patients with stage I-IV lung cancer by sing six publicly available data sets (1,689 patients, of whom 1,110 were diagnosed with non–small-cell lung) [1, 4, 11, 45, 75, 5]. IPRO showed a Concordance-index (C-index) of 0.72 for 1-year, 0.70 for 2-year, and 0.68 for 5-year mortality, performing a five-fold cross-validation. In [48], Haamburger et al. showed that by simplifying survival analysis to median survival classification, convolutional neural networks can be trained with small batch sizes and learn features that concatenated with radiomics features predict survival equally well as end-to-end hazard prediction networks. They obtained a C-index of 0.623 over 100 random splits, where for each split, 60%, 15% and 25% of the data was used for training, validation and testing, respectively. Direct hazard predictions from a neural network with radiomics features (multi-modal) and without were less precise with C-index of 0.613 and 0.585 respectively. Although these studies have demonstrated promising results in both classification and regression tasks, there are several notable limitations. A major limitation is represented by the high dimensionality of CT data and model complexity. The use of 3D CNNs in survival prediction models, as in studies like Zheng et al. [171] and Torres et al. [150], involves processing vast amounts of image data. Although 3D CNNs can capture volumetric details, the high dimensionality leads to increased computational costs and the risk of overfitting, particularly when datasets are limited in size. We propose a method that integrates a 2D convolutional neural network (CNN) with a soft attention mechanism to create a comprehensive representation of the 3D volume for the 2-year OS prognosis task. This approach aims to leverage the benefits of automatic feature extraction from images while focusing on relevant slices within the 3D images, which is achieved through the attention mechanism. Using 2D CNNs with soft attention reduces the complexity of the model

compared to 3D CNNs, which can lead to easier training and better generalization, especially when data is limited. Another limitation consists in the underutilization of temporal dynamics in survival prediction since these state-of-the-art papers primarily focus on fixed time points (e.g., 2-year survival), not accounting for the dynamic nature of survival, where the risk of mortality changes over time. We address this by processing the enhanced representation through the DeepHit risk-assessment network [86], as introduced in Section 3.3.1 for handling censored data and optimizing dynamic survival predictions over multiple time points. To assess the effectiveness of the proposed approach, we employ the time-dependent C-index (C^{td}-index), which is particularly suitable for OS regression as it accounts for the time to the event. This index offers a more comprehensive evaluation of model performance over the entire time horizon, rather than limiting the assessment to a single endpoint. Consequently, it provides deeper insights into the model's ability to predict survival risks at different intervals, ultimately leading to more robust and accurate survival predictions.

4.2 Materials and Methods

4.2.1 Datasets (LUNG1 and CLARO)

Two datasets were used for experiments: the publicly available LUNG1 dataset [4] and a proprietary dataset, hereafter referred to as CLARO. The LUNG1 dataset consists of 422 patients with stages I-III NSCLC. We excluded seven patients from the 422 due to missing CT slices in the complete volume (i.e. LUNG1-014, LUNG1-021, LUNG1-085, LUNG1-095, LUNG1-128, LUNG1-194, LUNG1-246). The CLARO dataset consists of 119 patients with stage III NSCLC, who received concurrent radiochemotherapy at the radiotherapy department of Campus Bio-Medico University, and has already been utilized in our previous studies [151, 152]. The 2-year OS distributions of the two datasets are depicted in Table 4.1. For both datasets, we employed only the pretreatment CT scans, preprocessed as in Section 4.2.

4.2.2 CT Preprocessing and Slice Selection Framework

To ensure comparability across patients and mitigate heterogeneity in acquisition protocols, a structured preprocessing and slice selection pipeline was implemented. The objective was to reduce the dimensionality of CT volumes while preserving the most prognostically relevant regions. The pipeline of a single CT scan is depicted in Figure 4.1.

In the initial step, we extracted a 3D array with voxel intensity values represented as Hounsfield Units from Digital Imaging and Communications in Medicine (DICOM) data

Table 4.1: 2-Year OS distributions for LUNG1 and CLARO datasets.

Dataset	Outcome	A-priori Probability
LUNG1	0 (Survivor)	0.40
	1 (Non-Survivor)	0.60
CLARO	0 (Survivor)	0.57
	1 (Non-Survivor)	0.43

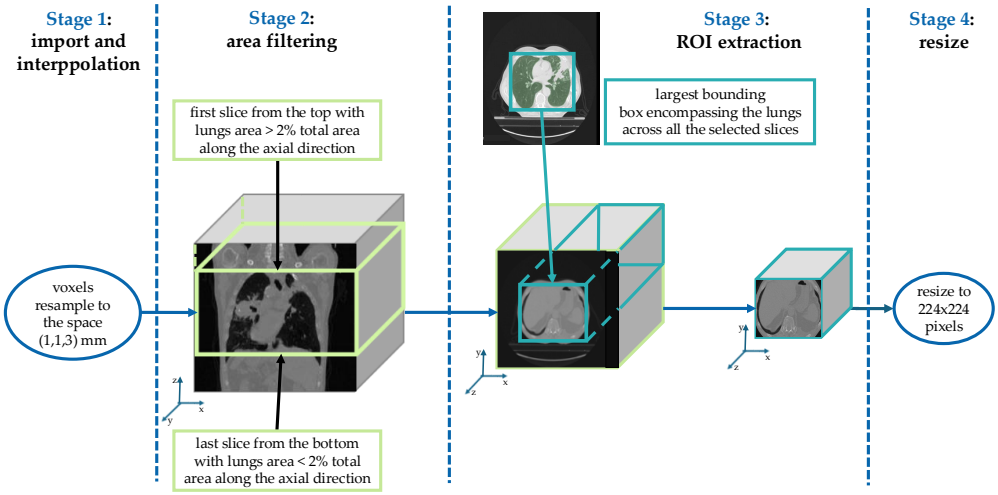

Figure 4.1: Preprocessing Pipeline.

and resampled the image voxels to a resolution of (1, 1, 3) mm to standardize the spatial dimensions of the CT scan. This resolution strikes a balance between spatial detail and computational efficiency. A voxel size of 1 mm in the x and y directions ensures high spatial resolution in the axial plane, which is essential for accurately visualizing fine details of lung tumors and lesions. Meanwhile, the 3 mm resolution in the z direction reduces the number of slices needed to cover the entire volume, thereby lowering computational demands without significantly sacrificing diagnostic quality.

Next, we applied an area filtering criterion to isolate a specific volume of the CT scan. This step utilized a U-net model, trained specifically for lung segmentation [57], to segment each individual slice and extract the right and left lungs. Slices with lung areas less than 2% of the total slice area were excluded to remove apical or basal slices containing minimal lung tissue, while the first slice below this threshold was retained to ensure complete coverage. This threshold was empirically chosen to balance prognostically relevant regions and

computational efficiency, as processing entire CT volumes with full peripheral slices would significantly increase GPU memory usage.

In the third step, we focused on extracting the region of interest (ROI). Once all slices were segmented, we extracted the ROI by identifying the largest bounding box that encompassed the lungs across all the selected slices and applied it to the entire volume.

Finally, we resized each slice to 224×224 pixels to standardize the input size for subsequent processing stages, thus enabling the use of pre-trained models on ImageNet.

4.2.3 Learning Prognostic Features from CT Scans

The proposed architecture, illustrated in Figure 4.2, combines slice-level feature representations through an attention mechanism to derive patient-level prognostic embeddings. The design exploits the efficiency of 2D convolutional neural networks (CNNs) while integrating inter-slice dependencies via attention, striking a balance between computational tractability and volumetric contextualization.

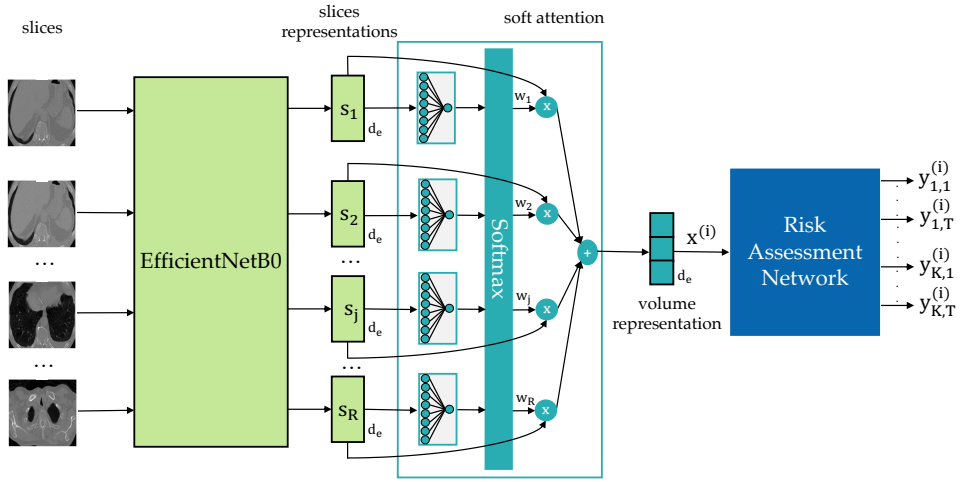

Figure 4.2: Overview of the proposed pipeline for slice-level feature extraction and attention-based aggregation.

CNN Encoder for Slice-Level Feature Extraction

Each axial slice of the CT scan was processed independently by a 2D encoder network. We selected the **EfficientNetB0** as our backbone due to its numerous advantages. The primary benefit is its exceptional efficiency. EfficientNet-B0 consists of a low number of parameters and FLOPS (Floating Point Operations per Second), detailed in Figure 4.3. This ensures that we can obtain high-quality features while minimizing the number of parameters,

reducing the risk of overfitting on small datasets. Moreover, EfficientNet-B0 is a versatile network. Its efficient feature extraction capabilities make it suitable for a wide range of computer vision tasks, enhancing its applicability across different projects [122, 102, 147, 9]. This architectural efficiency makes it particularly suitable for data-limited medical imaging scenarios.

The encoder was initialized with weights pretrained on ImageNet, enabling the transfer of low-level feature extraction capabilities (e.g., edge and texture detection) to the medical domain. To reduce overfitting risks, only the final layers of the backbone were fine-tuned on the survival prediction task. For each slice j belonging to patient i, the encoder produced a dense feature vector $\mathbf{s}_j^{(i)} \in \mathbb{R}^{d_e}$, capturing local prognostic cues such as tumor boundaries, parenchymal alterations, and lymph node involvement.

Inter-Slice Attention for Aggregating Prognostic Cues

Although naïve strategies such as average pooling treat all slices equally, they ignore the fact that only a subset of slices carries meaningful prognostic information (e.g., tumor-bearing regions). To address this, we employed a **soft attention mechanism** that dynamically assigns weights to slice embeddings according to their relevance to survival prediction.

For each slice representation $\mathbf{s}_j^{(i)}$, attention weights were obtained through a fully connected transformation followed by a softmax normalization:

$$w_j^{(i)} = \frac{\exp\left(\mathbf{W}\mathbf{s}_j^{(i)} + b\right)}{\sum_{k=1}^{R} \exp\left(\mathbf{W}\mathbf{s}_k^{(i)} + b\right)}, \quad j = 1, \ldots, R,$$

where $\mathbf{W} \in \mathbb{R}^{1 \times d_e}$ and $b \in \mathbb{R}$ are learnable parameters, and R denotes the number of slices for patient i.

The aggregated patient-level embedding $\mathbf{x}^{(i)} \in \mathbb{R}^{d_e}$ was then computed as the weighted sum of slice representations:

$$\mathbf{x}^{(i)} = \sum_{j=1}^{R} w_j^{(i)} \, \mathbf{s}_j^{(i)}. \tag{4.1}$$

This formulation enables the model to prioritize slices containing critical prognostic cues, while down-weighting redundant or non-informative regions. By generating a compact yet expressive representation of the entire CT volume, the method leverages the strengths of 2D CNNs combined with attention, thus avoiding the high computational burden of full 3D CNNs.

Overall, this approach reduces model complexity, accelerates training, and improves gen-

eralization in scenarios with limited data availability, all while retaining the volumetric context necessary for accurate survival prediction.

Risk Assessment Network

The representation $\mathbf{x}^{(i)}$ is fed into the DeepHit risk assessment network [86], detailed in Section 3.3.1. This network learns the distribution of survival times directly, without assumptions on the underlying stochastic process, and handles right-censored data through its specialized loss function. Here, we apply it to CT-derived features, maintaining the architecture with a shared sub-network and one cause-specific sub-network (for $k = 1$, patient death). The output is a probability distribution over discrete time points up to a 2-year horizon, enabling estimation of the cumulative incidence function (CIF) as in Equation 3.3.

For implementation details, including the loss function $L = \alpha L_1 + \beta L_2$ that jointly accounts for log-likelihood and ranking objectives, we refer to Section 3.3.1. Reusing the same prediction architecture across different data modalities promotes methodological consistency and comparability of results, allowing performance differences to be attributed to the input representations rather than to changes in the survival modeling framework, while also avoiding unnecessary redundancy in the methodological description.

4.3 Experimental Setup

4.3.1 Comparative Analysis

To demonstrate the benefits and strengths of our methodology, we compared our model against several competitors, each selected and configured to provide a clear and rigorous comparison. Below, we outline the key questions we aim to address through these comparisons, along with a detailed description of the experimental setups for each competing method.

Question 1: How does our method compare to 3D networks? To assess the effectiveness of our approach for constructing a robust representation of 3D volumes for the 2-year OS risk prediction task, we carried out a comparative evaluation by replacing our approach with competitive 3D baselines: an 18-layer 3D ResNet (ResNet3D_18) [153] pre-trained on the Kinetics dataset [73], a 121-layer DenseNet3D (DenseNet3D_121), and a Medical Slice Transformer (MST) [108] leveraging DINOv2 [114]. We evaluated both the 3D networks and our approach by employing different layer-freezing strategies. We evaluated two training strategies. First, we froze half of the layers in both the ResNet3D and the 2D backbone

(EfficientNetB0) of our approach. Next, we opted to freeze all layers except the last one (feature extractor), based on the hypothesis that reducing the number of parameters to optimize would enhance generalization with a limited dataset. This systematic evaluation allowed us to understand the impact of various layer-freezing configurations on the performance and efficiency of the models.

Question 2: How effective is the soft attention mechanism? We aimed to demonstrate the efficacy of the attention mechanism by comparing it with two alternative approaches: a simple average of the slices representations generated by the backbone, and a self-attention mechanism that incorporates a class token (CTk) summarizing the entire volume before being passed to the risk-assessment network. This set of experiments was conducted by freezing all backbone layers except the last one, as this configuration yielded higher performance in previous experiments.

Question 3: Is our approach adaptable to different 2D backbones? To illustrate the adaptability of our approach with various 2D backbones, we conducted another set of experiments using as backbone widely recognized architectures, including VGG16 [137], ResNet50 [81], AlexNet [81], Vision Transformer (ViT) [6], and MedViT_small [103]. These architectures were selected to represent different foundational paradigms, namely, convolutional networks (e.g., VGG16, ResNet50, AlexNet) and attention-based models (e.g., Vision Transformer, MedViT), to evaluate the versatility of our approach across both convolutional and attentional frameworks. This set of experiments also was conducted by freezing all backbone layers except the last one.

Question 4: What is the impact of fine-tuning? To underscore the significance of transfer learning, particularly when dealing with limited data, we conducted two experiments using the CLARO dataset. One experiment involved training without fine-tuning on LUNG1, whilst the other included fine-tuning on LUNG1.

Question 5: What is the computational complexity of the proposed method? To demonstrate the efficiency and cost-effectiveness of our approach, we conducted a computational complexity analysis of the various neural network configurations by considering the mean number of slices in the dataset, i.e. 72 and includes the number of parameters and the FLOPS for each configuration, providing insight into the computational demands of each model.

Table 4.2: Hyperparameters of the risk-assessment network.

Hyperparameter		**Value**
# hidden layers for both SN and CSNs		5
# neurons per hidden layer		100
Dropout rate		0.2
Activation function		ReLU
Loss function $\mathcal{L}$	α	0.5
	β	0.5

4.3.2 Performance metric

We evaluated the prognostic performance of our model using the time-dependent concordance index (C^{td}-index), as defined and detailed in Section 3.5. The C^{td}-index measures the model's ability to correctly rank pairs of patients based on survival times, accounting for censoring. The significance of this discriminative index lies in its independence from a single fixed time. This characteristic renders it well-suited for situations where the impact of covariates on survival varies over time. In other words, this index is particularly valuable when risks exhibit non-proportional behavior throughout the observation period.

4.3.3 Experimental settings

We implemented all considered methods in PyTorch [121] and trained them on an NVIDIA A100 40GB GPU. The training data augmentation strategy involved randomly flipping and rotating within a range of 20° each 2D slice with a probability of 0.2. During the training and evaluation of the risk-assessment network, we employed 10-fold cross-validation. For each fold, the 90% portion allocated for training was further split using a 90-10 ratio (90% for training and 10% for validation). All tested networks were trained with AdamW optimizer with base learning rate 1×10^{-4}, weight decay 1×10^{-2} and batch size 4 for 100 epochs. The best-performing model configuration was selected based on the highest time-dependent concordance index (C^{td}-index) achieved on the validation set. The loss weights and the other DeepHit hyperparameters were not optimized but were kept fixed at the values listed in Table 4.2. For the experiments on the CLARO dataset, we selected the best model configuration for each fold based on performance on the LUNG1 validation set. We then

fine-tuned this configuration using the CLARO training set.

4.4 Results and Discussion

We conducted a series of experiments to thoroughly evaluate the performance and flexibility of our proposed approach. First, we aimed to compare it against 3D baselines in survival prediction tasks, particularly focusing on the impact of employing a soft attention mechanism. Additionally, we tested the adaptability of our method by integrating alternative 2D backbones, such as ResNet50 and ViT, to assess whether it can maintain strong performance across diverse architectures. Moreover, we explored the significance of domain-specific transfer learning by fine-tuning models pre-trained on the LUNG1 dataset to investigate its impact on the smaller CLARO dataset and performed a detailed computational complexity analysis to determine the efficiency of our approach and its viability for survival prediction tasks in resource-limited environments. In the following, we present the results addressing the questions raised in Section 4.3.2.

Question 1: How does our method compare to 3D networks? Table 4.3 summarizes the results of the initial set of experiments, where our approach was compared with competitive 3D baselines: an 18-layer ResNet3D, a 121-layer DenseNet3D, and a Medical Slice Transformer. As shown in the table, our fusion approach, which combines EfficientNetB0 with a soft attention mechanism, consistently outperforms the 3D models under both layer-freezing strategies. This demonstrates the superior capacity of our method to create a consistent volume representation suitable for the OS prediction task, which is crucial for the risk-assessment network. Notably, the best performance is achieved with the EfficientNetB0 as the feature extractor, indicating that optimizing fewer parameters leads to better generalization when dealing with limited data availability. Given that we only one measurement per fold in the cross-validation, classical significance tests such as the paired t-test or Wilcoxon signed-rank test are not reliable, as their statistical power is limited in such low-sample settings and their distributional assumptions may not hold. To address this, we adopted a permutation-based paired t-test. In this approach, the observed difference in C^{td}-index between two models is compared against a null distribution obtained by repeatedly (5000 times) randomly swapping (with probability 0.5) the predictions of the two models at the patient level. The p-value is then computed as the proportion of permuted differences as extreme as the observed one. This non-parametric procedure makes no assumptions about the underlying distribution and is therefore more robust in our context of limited sample size. The results show that our proposed method (EfficientNetB0 + Soft Attention)

Table 4.3: Comparison between our approach and ResNet3D_18. Results of C^{td}-index over 10-fold cross-validation.

Approach	**Training Strategy**	C^{td}**-index (mean ± std)**
ResNet3D_18	half layers freezed	0.528 ± 0.069
DenseNet3D_121	half layers freezed	0.5331 ± 0.038
MST	half layers freezed	0.4915 ± 0.066
EfficientNetB0 + Soft Attention	half layers freezed	0.572 ± 0.064
ResNet3D_18	feature extractor	0.507 ± 0.074
DenseNet3D_121	feature extractor	0.5281 ± 0.056
MST	feature extractor	0.504 ± 0.051
EfficientNetB0 + Soft Attention	feature extractor	$\mathbf{0.584 \pm 0.054}$

achieves statistically significant improvements compared to all 3D baselines (ResNet3D_18, DenseNet3D_121, and MST), with $p < 0.05$ in all cases. This confirms that the observed performance gains are unlikely to be due to chance. The only non-significant comparison ($p = 0.4162$) arises when comparing our approach trained with the *feature extractor* strategy against the same model trained with *half of the layers frozen*. This result highlights that the advantage comes not only from the architecture itself but also from the training strategy: using EfficientNetB0 as a frozen feature extractor reduces the number of trainable parameters and thus improves generalization under limited data availability.

Question 2: How effective is the soft attention mechanism? We compared the soft attention approach with a simple average of the slices' representations generated by the backbone, and a self-attention mechanism incorporating a class token (CTk) summarizing the entire volume. We selected in all the experiments the EfficientNetB0 as feature extractor since it was the experiment with the best performance in the previous analysis in paragraph 4.4. Table 4.4 shows the results of this comparison study.

The table demonstrates that the fusion approach using the soft attention mechanism yields superior performance compared to both the average and self-attention approaches. The superior performance of the soft attention mechanism over the Average Pooling approach can be attributed to its ability to weigh the most important slices within the volume.

Table 4.4: Comparison between the soft attention mechanism with the Average Pooling and Self Attention approaches. All backbone layers are frozen, except for the last one (feature extractor). Results of C^{td}-index over 10-fold cross-validation.

Approach	**Aggregation Mechanism**	**C^{td}-index (mean ± std)**
EfficientNetB0	Average Pooling	0.570 ± 0.033
EfficientNetB0	Self Attention + CTk	0.579 ± 0.079
EfficientNetB0	**Soft Attention**	$\mathbf{0.584 \pm 0.054}$

This weighting allows the risk-assessment network to focus on the most relevant portions of the volume, which are more critical for OS prediction. When compared to the self-attention mechanism, the superior performance of soft attention may be due to its better generalization capability and the optimization of fewer parameters, which together can lead to more effective learning and improved performance, especially when dealing with limited data, as shown in the previous experiments. To assess the statistical significance of these differences, we employed a permutation-based paired t-test across the 10 cross-validation folds. The results confirmed that the improvements of soft attention over average pooling ($p = 0.5450$) and over self-attention ($p = 0.3240$) were not statistically significant. Nevertheless, the proposed method consistently achieved the highest mean C^{td}-index across folds, supporting its robustness. Beyond the numerical results, soft attention provides a more flexible aggregation strategy by weighting slices according to their relevance, providing interpretable insights into which regions contribute most to the prediction. In addition, compared to the self-attention mechanism, soft attention requires fewer parameters and is computationally more efficient, making it more suitable in scenarios with limited data and for practical deployment.

Question 3: Is our approach adaptable to different 2D backbones? Table 4.5 reports the results of these experiments. EfficientNetB0's superior performance with respect to other 2D backbones stems from its compound scaling method, which balances network depth, width, and resolution, optimizing its architecture for both accuracy and efficiency. Additionally, its streamlined architecture, characterized by a reduced number of parameters, makes it particularly well-suited for data-limited scenarios, where it effectively minimizes the risk of overfitting while preserving high computational efficiency. As evidenced by the results, our fusion approach achieves higher performance with different backbones compared to the 3D approach. This demonstrates that the mechanism of combining 2D slice representations to construct a 3D volume representation is more effective for OS prediction, particularly

Table 4.5: Comparison between the EfficientNetB0 with other 2D backbones in combination with the Soft Attention mechanism. All approaches are compared with the 3D approach. All backbone layers are frozen, except for the last one (feature extractor). Results of C^{td}-index over 10-fold cross-validation.

Approach	**C^{td}-index (mean ± std)**
ResNet3D_18	0.507 ± 0.074
ResNet50 + Soft Attention	0.564 ± 0.067
Vgg16 + Soft Attention	0.524 ± 0.063
AlexNet + Soft Attention	0.550 ± 0.0510
ViT_b_16 + Soft Attention	0.546 ± 0.043
MedViT_small + Soft Attention	0.5482 ± 0.056
EfficientNetB0 + Soft Attention	$\mathbf{0.584 \pm 0.054}$

when working with limited data. The enhanced performance underscores the advantage of leveraging 2D networks over directly using 3D networks in this context.

Question 4: What is the impact of fine-tuning? The results of these experiments are presented in Table 4.6. Since the CLARO dataset is extremely small, we did not report the standard deviation. We had to aggregate the test folds to obtain a single measurement, making it impossible to calculate variability across folds. The results clearly indicate that domain-specific transfer learning from the LUNG1 dataset leads to better performance on the CLARO dataset compared to fine-tuning. This improvement can be attributed to the limited data available in the CLARO dataset, which makes it challenging to fine-tune a model effectively. In contrast, leveraging the knowledge learned from the LUNG1 dataset provides a substantial boost, as the pre-trained model can utilize previously acquired patterns and features, thereby enhancing the performance on the smaller CLARO dataset. This underscores the value of domain-specific transfer learning in scenarios with constrained data availability.

To further demonstrate the robustness of our approach, we repeated the experiments five additional times and performed statistical testing. With fine-tuning on LUNG1, the model achieved an average performance of 0.577 ± 0.021, whereas training from scratch resulted in 0.503 ± 0.017.

Table 4.6: Results of C^{td}-index on CLARO. All backbone layers are frozen, except for the last one (feature extractor).

Approach	**Fine-tuning on LUNG1**	C^{td}**-index**
EfficientNetB0 + Soft Attention	No	0.503
EfficientNetB0 + Soft Attention	**Yes**	**0.579**

To compare the two models we conducted a paired statistical test. Each model was evaluated over the six repeated runs, with each run producing a single aggregated performance metric from a 10-fold cross-validation. The pairing is feasible because the same data splits and experimental conditions were used for both models in each run, ensuring that performance differences reflect the model initialization strategy rather than variability in the data.

Given the small number of paired samples ($n = 6$) and the unknown distribution of performance differences, we used the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test. This test assesses whether the median difference between paired observations is significantly different from zero without assuming normality. Additionally, we report the effect size (r) to quantify the magnitude of the observed difference.

The Wilcoxon signed-rank test yielded a statistic of 0.0, with a p-value of 0.03125 and a large effect size (r = 0.899)

These results indicate a statistically significant difference between the two models ($p < 0.05$), with the pretrained model consistently outperforming that trained from scratch. The large effect size ($r \approx 0.9$) further confirms that this difference is not only statistically significant but also practically meaningful.

In conclusion, pretraining the EfficientNetB0-based architecture provides a clear advantage over training from scratch, and the Wilcoxon signed-rank test robustly supports this finding despite the limited number of repeated runs.

This improvement can be attributed to the limited data available in the CLARO dataset, which makes fine-tuning a model from scratch challenging. In contrast, leveraging the knowledge learned from the LUNG1 dataset provides a substantial boost, as the pretrained model can exploit previously acquired patterns and features, thereby enhancing performance on the smaller CLARO dataset. These findings underscore the value of domain-specific transfer learning in scenarios with constrained data availability.

In addition, to further assess the generalizability of our approach, we directly evaluated the best-performing model selected from the 10-fold cross-validation on the external LUNG1

dataset without any fine-tuning. The model achieved a C^{td}-index of 0.5608, indicating that the learned representations retain prognostic relevance beyond the training cohort, despite the absence of dataset-specific adaptation.

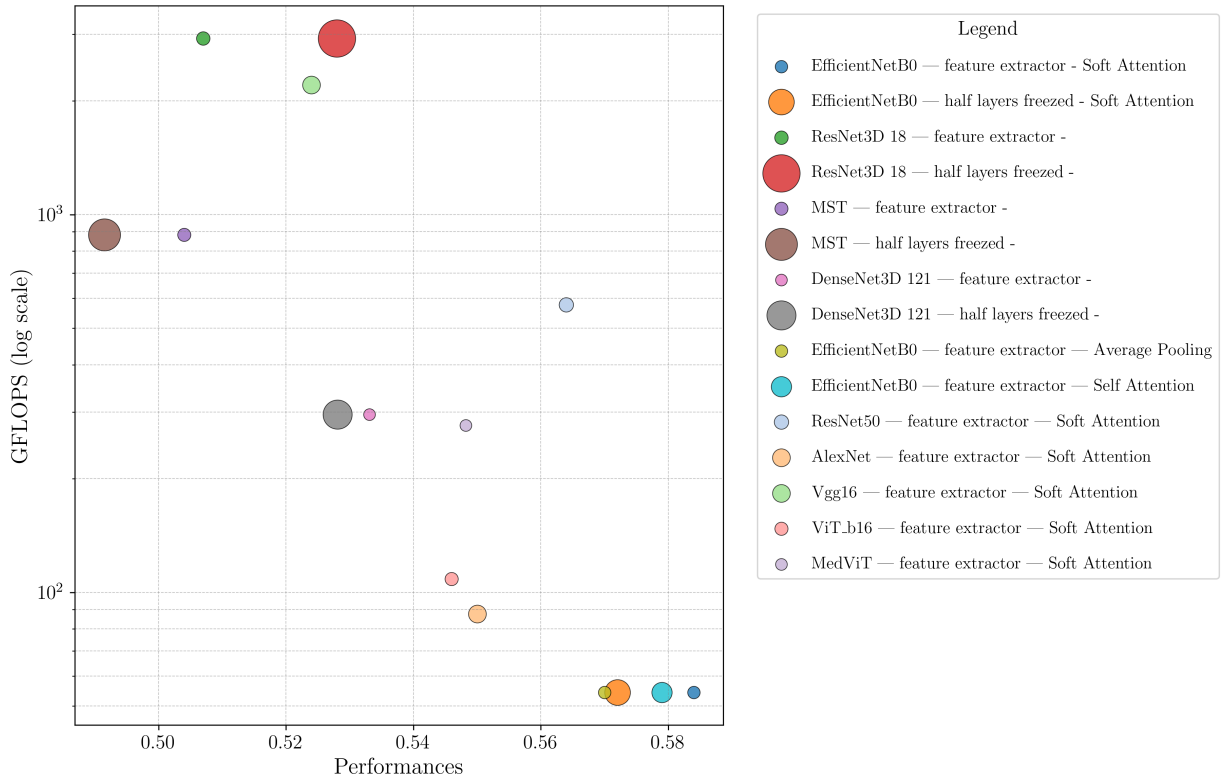

Figure 4.3: Computational complexity analysis of the experimented models. The x-axis represents the model's performance, while the y-axis shows the number of giga floating-point operations per second (GFLOPS) required for inference. The size of each point is proportional to the number of parameters of the corresponding model, providing a visual representation of the trade-off between model accuracy, computational cost, and model size.

Question 5: What is the computational complexity of the proposed method? The computational complexity analysis of the various model configurations is detailed in Figure 4.3. The figure is a bubble chart comparing various model configurations based on their performances (x-axis) and GFLOPS (y-axis, log scale). Each bubble represents a specific model, with its size indicating the number of parameters. The analysis highlights the trade-offs between performance and computational cost across different model architectures and layer-freezing strategies. Notably, EfficientNetB0 variants are characterized by lower parameter counts and computational costs, making them a suitable choice for resource-constrained environments with high performance. In the versions of EfficientNet with the lowest costs,

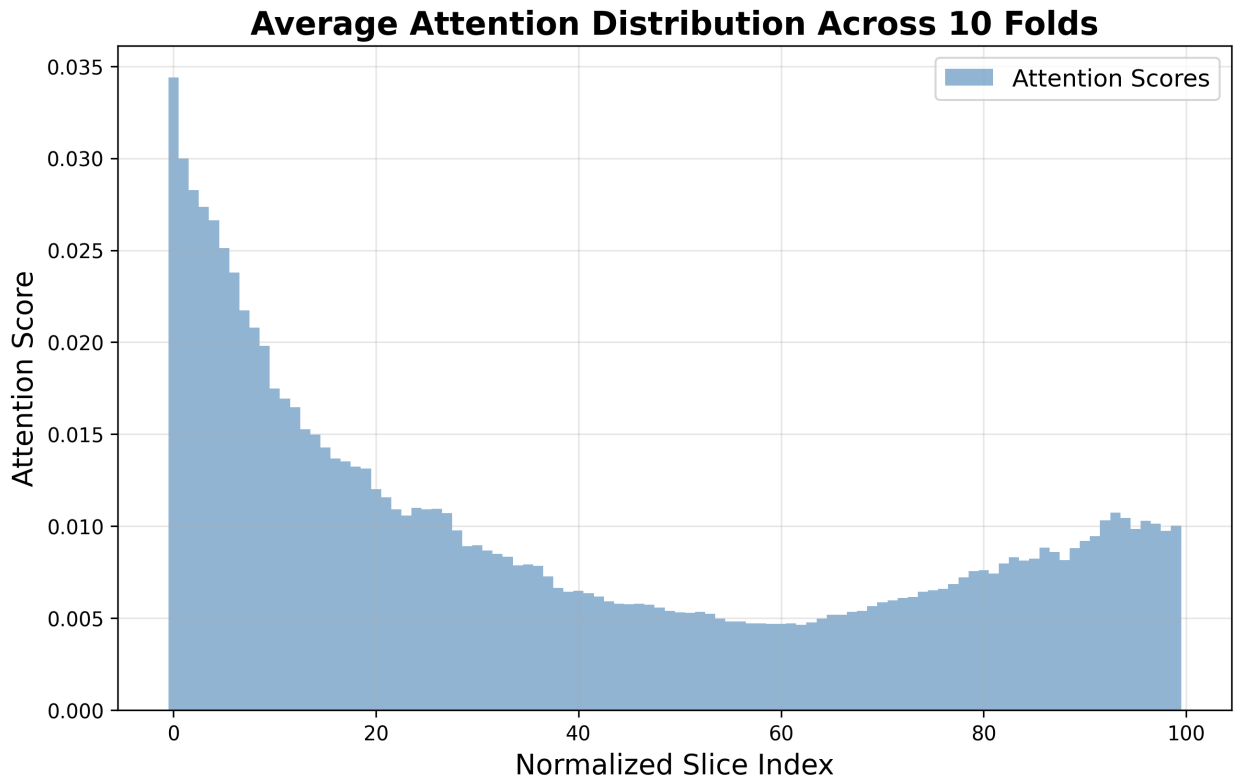

Figure 4.4: Average attention weight distributions generate by our model.

the soft attention mechanism is utilized to further enhance efficiency. In comparison, other 2D backbones demonstrate inferior performance and higher computational costs relative to EfficientNetB0, underscoring the crucial role of parameter efficiency in data-limited scenarios. On the other hand, 3D networks yields significantly higher computational costs and parameter counts, leading to lower efficiency and performance.

4.5 Interpretability and Clinical Insight

Interpretability remains a fundamental challenge in deep learning, particularly in medical applications: understanding why a model produces a specific prediction is critical for clinical trust and adoption. We address this challenge by leveraging attention weights derived from a soft attention mechanism to identify the CT scan slices that the model considered most informative for prognostic prediction. By highlighting these slices, we provide a window into the model's decision-making process, thereby bridging the gap between algorithmic output and clinical understanding.

Our model was trained to assign attention weights to individual slices within a 3D CT volume, quantifying their relative contribution to the final prognostic outcome. For the

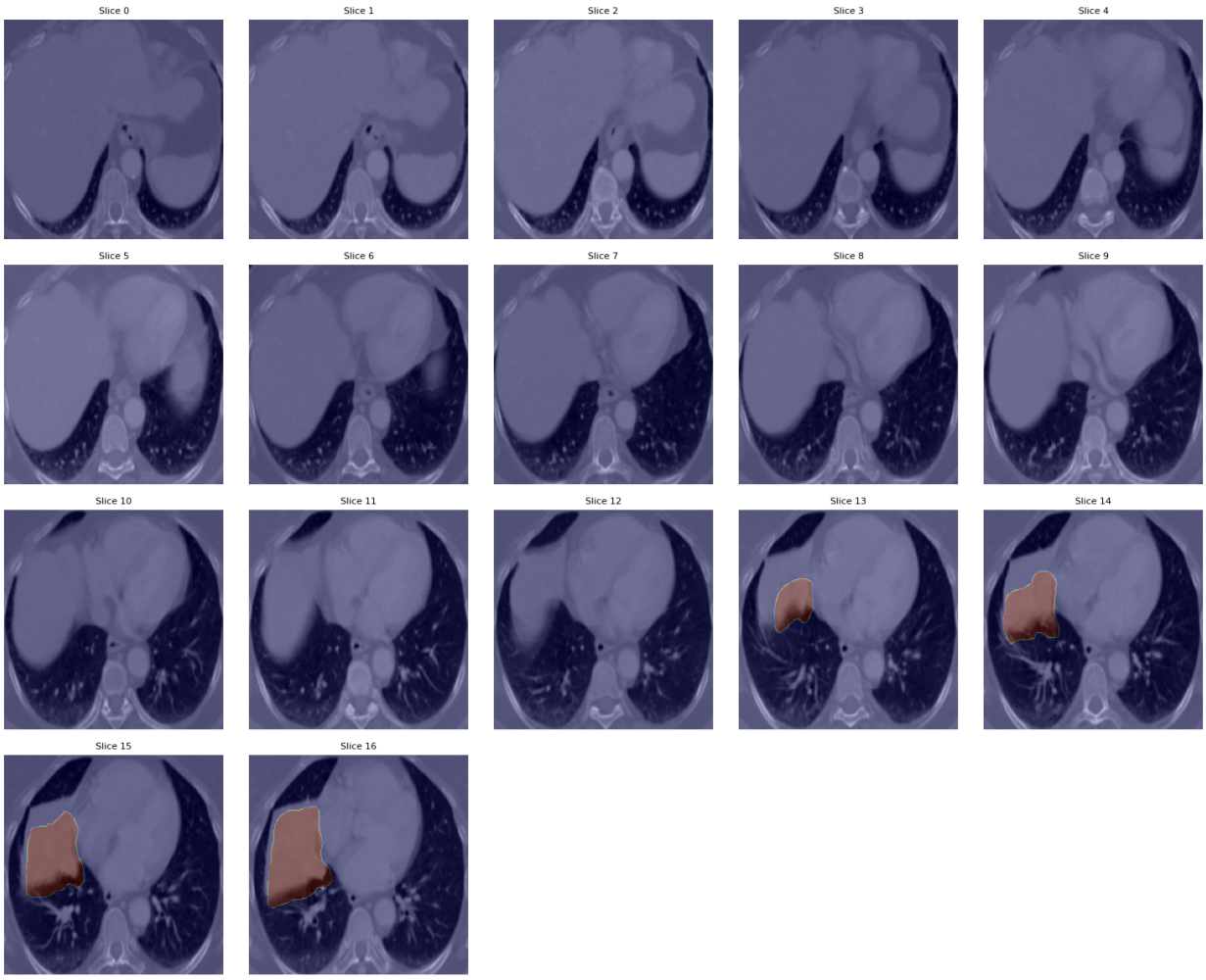

Figure 4.5: Example of attention weight distribution for patient LUNG1-040 (stage III NSCLC, total slices = 57). Slices in the 50th percentile of the mean attention distribution are highlighted, showing a concentration of attention in the basal lung regions where the tumor is located, consistently with the adanced stage of the patient.

interpretability analysis, we relied on the models trained in the 10 cross-validation folds. For each fold, attention weights were computed on the corresponding test set. Since the number of slices varies between patients, we normalized slice positions by calculating relative slice indices. Subsequently, we averaged the attention weights across all patients and folds to obtain a robust, mean attention distribution.

Analysis of this aggregated distribution revealed a consistent and clinically meaningful pattern, as illustrated in Figure 4.4. Specifically, basal lung regions consistently received higher attention weights from the model. This observation aligns closely with clinical evidence: tumors located in the basal regions of the lung are often associated with worse outcomes in non-small cell lung cancer (NSCLC). Basal tumor location has been linked to

several prognostic factors, including:

- **Advanced stage and aggressiveness:** Tumors in basal regions are frequently detected at later stages and may exhibit higher invasive potential due to delayed symptom onset and the complexity of anatomical positioning [87].
- **Proximity to vital structures:** Basal tumors are anatomically positioned near critical organs such as the heart, great vessels, and diaphragm, which can complicate surgical interventions and directly impact prognosis.
- **Increased symptom burden:** These tumors often result in more severe respiratory impairment and cardiac complications, including pleural effusion and atelectasis, negatively affecting both survival and quality of life.

The correspondence between the model's attention focus and established clinical knowledge reinforces the biological plausibility of the learned features. This not only enhances interpretability but also increases confidence in the model's ability to capture clinically relevant patterns, which is essential for translational applications in precision oncology. By highlighting CT slices associated with known prognostic indicators, attention-based interpretability enables a more transparent and informative dialogue between AI predictions and clinician decision-making.

An illustrative example is shown in Figure 4.5, highlighting the slices in the 50th percentile of the mean attention weight distribution for patient LUNG1-040 (total slices = 57) with stage III NSCLC. The tumor is present in the lower (basal) slices, consistent with its advanced stage and basal position, which underscores the model's ability to capture clinically meaningful spatial patterns in the lung. Furthermore, the patient's observed survival time was 19 months, and the model assigns a corresponding probability of 0.61 for the occurrence of the event "death" at that time, demonstrating coherent alignment between model prediction and clinical outcome.

Chapter 5

Multimodal Clinical Data Fusion for Prognostic Modeling in Pulmonary Diseases

In previous chapters, this thesis investigated the enhancement of predictive models for NSCLC using single-modality data sources, specifically unstructured EHRs (Chapter 3) and CT scans (Chapter 4). Although multimodal models have the potential to outperform unimodal approaches by integrating complementary information, their development has been hindered by the lack of publicly available NSCLC datasets that offer sufficient sample size, complete modality coverage (CT imaging, structured and unstructured EHR), and prognostic labels.

To overcome this limitation, the present chapter shifts focus to the INSPECT dataset, a publicly available resource originally designed for pulmonary embolism (PE) prognosis. INSPECT includes CT pulmonary angiography (CTPA) scans, radiology reports, and structured EHR data, all aligned with mortality outcomes at 1-, 6-, and 12-month intervals. This dataset was selected due to the shared pulmonary context between PE and NSCLC: both conditions require thoracic imaging and clinical data integration for effective prognostic modeling. Although PE and NSCLC differ in etiology, the high incidence of PE in lung cancer patients, and its demonstrated association with increased morbidity and mortality [112, 177], support the use of PE as an informative surrogate to explore multimodal fusion strategies relevant to lung disease prognosis. The comprehensive modality coverage and prognostic task alignment of INSPECT make it a suitable testbed to extend the methodologies previously developed in this thesis, particularly NER (Chapter 3) and slice-level attention for CT imaging (Chapter 4), within a multimodal fusion framework.

This chapter presents a comparative study of multimodal data integration for PE prognosis, advancing the overarching goal of this thesis: improving predictive modeling in lung diseases. Beyond advancing PE prognosis, these findings offer transferable lessons for future NSCLC research, addressing current gaps in modality selection and fusion strategy optimization.

5.1 Motivation and Clinical Background

PE is a potentially life-threatening cardiovascular condition caused by obstruction of the pulmonary arteries, most commonly due to thromboembolic events. Despite improvements in diagnostic imaging and anticoagulant therapies, PE remains associated with substantial short- and medium-term mortality, highlighting the clinical importance of accurate risk stratification and prognostic assessment [12]. The prognostic impact of PE is even more pronounced in oncologic populations: recent studies have shown that incidental or unsuspected PE significantly worsens long-term survival in patients with cancer [112]. This association is particularly relevant for thoracic malignancies, where tumor-related hypercoagulability, treatment toxicity, and advanced disease burden jointly contribute to an increased thromboembolic risk. As emphasized in [177], PE represents a major determinant of morbidity and mortality in lung cancer patients, underscoring the need for reliable prognostic tools in this clinical setting.

Risk stratification in PE has traditionally relied on validated clinical scoring systems, most notably the Pulmonary Embolism Severity Index (PESI) and its simplified version (sPESI) [10, 69]. These scores combine demographic variables and readily available clinical features to estimate short-term mortality risk and are widely adopted in clinical practice. While effective for early decision-making, their reliance on a limited set of structured inputs restricts their ability to capture the multidimensional nature of PE, particularly information embedded in imaging studies, radiology reports, and longitudinal electronic health records (EHRs).

In recent years, machine learning (ML) and deep learning (DL) approaches have been increasingly explored to overcome these limitations. Unimodal models based on structured EHR data leverage demographic information, vital signs, laboratory values, and comorbidities using techniques such as logistic regression, gradient-boosted trees, and neural networks [39, 131]. Imaging-based approaches apply convolutional neural networks to CT pulmonary angiography (CTPA) to extract prognostically relevant features related to clot burden or cardiac strain [38]. In parallel, natural language processing (NLP) methods applied to radiology reports have evolved from rule-based systems to transformer-based architectures,

enabling the extraction of clinically meaningful semantic information for PE detection and, more recently, prognosis [128, 52, 113, 7]. Each of these unimodal approaches captures complementary aspects of PE but remains inherently limited when considered in isolation.

Multimodal deep learning (MDL) has therefore emerged as a promising paradigm to integrate heterogeneous clinical data sources and improve predictive performance. By jointly modeling imaging, textual, and structured EHR data, MDL approaches aim to exploit the complementary strengths of each modality. Fusion strategies can be broadly categorized into early, late, and intermediate fusion, each offering different trade-offs between flexibility, interpretability, and capacity to model cross-modal interactions [46]. Across multiple medical domains, multimodal integration has consistently demonstrated superior performance compared to unimodal models [144, 64, 68]. In the context of PE, studies combining imaging and clinical data have shown improved diagnostic and prognostic accuracy over traditional clinical scores and unimodal baselines [65, 16, 172].

Despite these advances, several critical limitations remain. First, most existing PE studies focus on diagnosis or short-term outcomes, with limited attention to clinically actionable medium-term prognostic horizons (e.g., 1-, 6-, and 12-month mortality). Second, the majority of multimodal approaches integrate only two modalities, typically CTPA and structured EHR, while underutilizing the rich qualitative and semi-structured information contained in radiology reports. Third, reproducibility and generalizability are hindered by the frequent use of proprietary datasets, limiting external validation and clinical translation.

The release of the INSPECT dataset [63] represents a major step toward addressing these challenges. INSPECT is the first large-scale, publicly available multimodal cohort that jointly provides CTPA images, radiology report impressions, and longitudinal structured EHR data, together with diagnostic and prognostic labels. While existing INSPECT baselines demonstrate the benefits of multimodal learning, they primarily adopt late fusion strategies and do not fully exploit the textual modality or systematically assess different fusion paradigms across multiple prognostic horizons.

Motivated by these gaps, this chapter aims to provide a comprehensive and reproducible evaluation of multimodal learning strategies for PE prognosis. Specifically, the main contributions are:

- **Systematic multimodal comparison:** A comprehensive evaluation of unimodal (CT, radiology reports, structured EHR) and multimodal models for PE prognosis across clinically relevant prediction horizons (1-, 6-, and 12-month mortality) using the INSPECT dataset.

- **Extended prognostic scope:** A focus on medium-term outcomes beyond diagnosis

and short-term risk assessment.

- **Comprehensive modality coverage:** An explicit assessment of the incremental value of imaging, textual, and structured EHR data and their combinations.
- **Fusion strategy analysis:** An empirical investigation of early, intermediate, and late fusion schemes to determine when and how modalities should be integrated.
- **Benchmarking and reproducibility:** The establishment of reproducible baselines to facilitate future multimodal PE research.

Together, these contributions provide both a methodological and practical foundation for the development of multimodal prognostic models in pulmonary embolism and related cardiovascular conditions.

5.2 Materials and Methods

5.2.1 The INSPECT Dataset

Experiments were conducted on the INSPECT dataset [63], the only publicly available multimodal resource for PE prognosis, which includes paired CTPA scans, radiology reports, and structured EHR data. The original dataset contains in total 23,248 CTPA studies from 19,402 patients, with some patients undergoing multiple CTPA exams. Our analysis is performed at the CTPA-study level. Accordingly, the dataset provides multimodal information collected around the "index CTPA study", defined as the CT angiography performed during the clinical visit in which pulmonary embolism was suspected or diagnosed.

Each sample includes the full 3D CTPA volume acquired at this index visit (acquisition details are reported in [63]) and the associated radiology report, which contains only the "Impression" section summarizing key PE-related findings. Full narrative reports are not provided.

The structured EHR data consist of the entire longitudinal clinical history available for each sample, including diagnoses (ICD codes), medications, procedures, laboratory tests, and vital signs. Importantly, INSPECT does not pre-align EHR events to the CTPA exam: each patient's EHR timeline is provided in full. In our study, we specifically selected only those events occurring prior to the index CTPA timestamp to ensure that prognostic modeling is based strictly on pre-diagnostic information, thereby avoiding any potential data leakage and reflecting clinically realistic prediction scenarios. We excluded samples for whom the

Time	Label	All	Train	Val	Test
1 m	pos.	1143	939 (5.45%)	53 (5.23 %)	151 (5.14%)
	neg.	20049	16305 (94.55%)	960 (94.77%)	2784 (94.86%)
	cens.	2056	-	-	-
6 m	pos.	2302	1891 (11.54%)	99 (10.21%)	312 (11.17%)
	neg.	17854	14501 (88.46%)	871 (89.79%)	2482 (88.83%)
	cens.	3092	-	-	-
12 m	pos.	2811	2303 (14.59%)	125 (13.51%)	383 (14.25%)
	neg.	16473	13369 (85.31%)	800 (86.49%)	2304 (85.75%)
	cens.	3964	-	-	-

Table 5.1: Distribution of prognostic labels across 1-, 6-, and 12-month mortality prediction tasks in the INSPECT dataset. Each case includes a CTPA scan, its corresponding radiology report, and structured EHR data. Patient-level splitting ensures that all cases from the same patient are assigned to the same set (train, validation, or test). Percentages in splits indicate the proportion of positive versus negative cases.

timestamp of the index CTPA study was missing, since this prevents consistent extraction of pre-exam EHR events.

INSPECT provides prognostic outcomes at 1-, 6-, and 12-months, enabling systematic evaluation across short- and medium-term horizons. Outcomes are annotated as positive (death within the target window), negative (survival), or censored (insufficient follow-up). In line with the original benchmarks, only positive and negative samples were considered, while censored cases were discarded, as shown in Table 5.1. This resulted in 21,192 CTPA studies for the 1-month mortality task, 20,156 for the 6-month task, and 19,284 for the 12-month task. Across these horizons, each patient contributes slightly more than one study, with values ranging from a minimum of 1 to a maximum of 13 CTPA studies per patient across all prognostic tasks.

To ensure reproducibility and prevent information leakage, particularly because individual patients may undergo multiple CTPA exams, the official patient-level INSPECT splits were adopted, assigning all CTPA studies from a given patient exclusively to a single set (training, validation, or test).

5.2.2 Data Preprocessing

Each modality in INSPECT required a tailored preprocessing pipeline to ensure comparability across samples and suitability for downstream modeling.

CTPA scans. For CTPA preprocessing we follow the same pipeline described in [63] . Each CTPA exam was preprocessed by extracting the pixel data from the original DICOM

format. Raw intensities were converted into Hounsfield Units (HU). To reduce computational cost, all slices were resized to 256×256 pixels and subsequently cropped around the lung region, resulting in a final resolution of 224×224.

To emphasize clinically relevant structures, three standard viewing windows were applied, each optimized for a specific diagnostic perspective: (i) lung parenchyma, (ii) pulmonary embolism, and (iii) mediastinum. The corresponding windowed images were stacked to form a three-channel representation of each slice ($224 \times 224 \times 3$). Finally, pixel intensities were normalized to zero mean and unit variance using ImageNet statistics (mean and standard deviation), ensuring compatibility with pretrained convolutional backbones ([34]).

Radiology reports. Free-text radiology reports were segmented into sentences using a custom algorithm designed to preserve clinically meaningful structures, including numbered findings (e.g., "1. PROBABLE"). The algorithm temporarily replaces periods in numbered lists (e.g., "1.") to prevent incorrect sentence splits, then applies standard sentence-ending punctuation rules by identifying spaces following ".", "?", or "!". Finally, the original numbering is restored. This procedure is summarized in Algorithm 1 and ensures that both the overall report structure and individual observations are accurately preserved. After sentence segmentation, each sentence was tokenized using the Clinical-Longformer tokenizer [94], ensuring compatibility with the subsequent transformer-based encoding.

Algorithm 1 Custom Sentence Splitter for Radiology Reports

procedure CUSTOMSENTENCESPLITTER(text)
    Input: raw radiology report text
    Output: list of sentences
    Replace all "\d+.\s" with "\1###"
    Split text at spaces following '.', '?', or '!'
    Restore numbering by replacing "###" with ". "
    Trim whitespace in each sentence
    Remove empty sentences
    return list of sentences
end procedure

Structured EHR data. Structured EHR events were represented using a count-based featurization strategy ([126]), producing a sparse patient-by-code matrix. Each entry in this matrix indicates the frequency with which a specific clinical code (e.g., diagnoses, medications, procedures, lab tests) occurs in a patient's record prior to the index "CTPA exam", ensuring that only information available at prediction time is used. This representation provides a compact summary of longitudinal patient history, retaining at the same time rich

information about prior comorbidities, interventions, and monitoring.

To incorporate EHR features into deep learning models, which typically require dense, low-dimensional representations, two complementary dimensionality reduction strategies were investigated .

The first approach relied on truncated singular value decomposition (TSVD) ([50]), projecting each patient vector into a 128-dimensional embedding that preserved approximately 99% of the variance of the original sparse matrix. This method retains most of the salient information, substantially reducing computational complexity, thus facilitating integration with the other modalities. We refer to this representation as the EHR-TSVD embedding.

The second approach employed a feed-forward autoencoder trained to reconstruct the original sparse matrix from a compact latent representation. The encoder output defines the EHR-AE embedding, a dense representation that summarizes each patient's longitudinal clinical history and can be seamlessly combined with imaging and text embeddings within multimodal architectures (see Section **??** for details).

5.2.3 Models

To systematically evaluate the contribution of different data sources and fusion strategies for PE prognosis, both unimodal and multimodal models were implemented. This section is organized into three main components.

First, the representation extraction process is described, in which each modality (structured EHR, radiology reports, and CTPA imaging) is encoded into a fixed-length sample-level representation using modality-specific encoders. This component establishes the common feature space shared by all subsequent models.

Next, the unimodal baselines are presented. Each modality is modeled independently using architectures tailored to its statistical and structural properties. These baselines serve two purposes: (i) they provide strong single-source predictors for PE prognosis, and (ii) they allow quantification of the standalone predictive value of each modality before integration into multimodal settings.

Finally, the multimodal fusion models are introduced, combining the representations extracted from different modalities through a set of fusion strategies. Both bimodal and trimodal configurations are detailed. The primary objective is to assess the added value of multimodal models over unimodal baselines. Among the multimodal strategies, bimodal combinations offer insight into which modality pairs contribute most to predictive performance and serve as ablation studies related to trimodal fusion.

Representation Extraction

For each data modality (structured EHR, radiology reports, and CTPA imaging) a fixed-length sample-level representations were extracted using modality-specific encoders designed to capture complementary aspects of a patient's clinical profile. The dimensionality of the resulting representations is summarized in Table 5.2.

Structured EHR data were converted into high-dimensional count vectors summarizing each sample's clinical events up to the time of the index CTPA exam. The resulting representation is a sparse vector of size $1 \times 74{,}303$ for each patient.

Table 5.2: Dimensionality of the intermediate representations extracted for each modality prior to aggregation.

Modality	**Encoder / Method**	**Representation Dimensionality**
Structured EHR	Count-based featurization	74,303
Radiology Report	Clinical-Longformer	[#Sentences, #Tokens, 768]
CTPA Imaging	ResNetV2-101	[#Slices, 6,145]

In radiology reports, the text was first split into sentences and tokens (see Section 5.2.2) and then encoded using Clinical-Longformer [94], a transformer model pretrained on large-scale clinical corpora (e.g., MIMIC-III [70]) and specifically optimized for long clinical documents. The model produced a contextualized embedding $\mathbf{t}_{i,j}$ of size 768 for each token j in each sentence i. This results in a tensor of shape [#sentences, #tokens, 768], where the "#" symbol denotes "number of", thus preserving the sentence-level structure prior to hierarchical aggregation.

For CTPA imaging, after the preprocessing described in Section 5.2.2 , each axial slice was processed using a ResNetV2-101 convolutional encoder pretrained with BigTransfer [79] and subsequently adapted to the pulmonary domain through fine-tuning on the RSPECT dataset [30], yielding slice-level representations $\mathbf{c}_i$ of size 6145 before slices-level aggregation. This results in a tensor of shape [#slices, 6145], where the "#" symbol denotes "number of". To ensure consistency across samples, the first 250 slices for scans exceeding this number were retained. This cutoff was chosen to standardize input dimensionality at the same time preserving the full thoracic region relevant for prognostic assessment, thus excluding abdominal sections that do not contribute with meaningful information and reducing computational overhead.

These modality-specific representations serve as the inputs for the unimodal baselines and the multimodal fusion models described in the following sections.

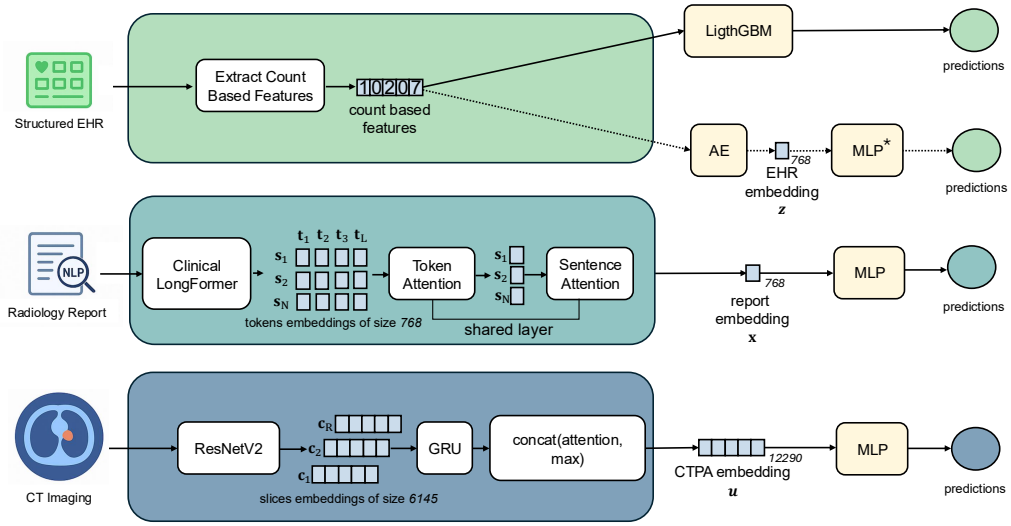

Figure 5.1: Overview of unimodal prognostic models. Each modality is processed independently: (i) CTPA slices are encoded with a ResNetV2-101 backbone followed by sequential modeling and pooling, (ii) radiology reports are embedded with Clinical-Longformer and aggregated via a two-levels hierarchical attention mechanism (Token Attention and Sentence Attention), and (iii) structured EHR data is featurized with count-based methods and modeled using gradient boosting or autoencoder.

Unimodal Models

Unimodal baselines are illustrated in Figure 5.1. Each modality was first modeled independently to establish strong single-source predictors before exploring multimodal fusion.

Structured EHR data For the Unimodal EHR (Structured EHR) experiments, a supervised autoencoder (EHR-AE) designed to learn compact, task-specific embeddings from structured, count-based EHR features was first introduced. Let $\mathbf{x} \in \mathbb{R}^d$ denote a sample-level EHR vector, with $d = 74303$. The encoder computes a latent representation

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x}), \qquad f_{\text{enc}} : \mathbb{R}^d \to \mathbb{R}^{768},$$

through a two-layer architecture with dimensions $d \to 512 \to 768$. The decoder mirrors this structure:

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z}), \qquad f_{\text{dec}} : \mathbb{R}^{768} \to \mathbb{R}^d.$$

A classification head predicts mortality:

$$\hat{y} = f_{\text{clf}}(\mathbf{z}), \qquad f_{\text{clf}} : \mathbb{R}^{768} \to [0, 1],$$

implemented as a two-layer MLP ($768 \to 128 \to 1$).

The model is trained end-to-end with a joint loss:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}) + \beta\, \mathcal{L}_{\text{BCE}}(y, \hat{y}),$$

with weights $\alpha = 0.5$ and $\beta = 1.0$, chosen to prioritize the mortality prediction task while still encouraging accurate reconstruction of the input. Here,

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \qquad \mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right].$$

This model learns dense, task-adaptive embeddings from structured EHR data, which can be directly used for downstream multimodal fusion.

For comparison, the gradient boosting EHR model from the INSPECT framework [63] was also reproduced. Let $f_{\text{GBM}}(\mathbf{x})$ denote the model prediction, defined as an ensemble of T regression trees:

$$f_{\text{GBM}}(\mathbf{x}) = \sum_{t=1}^{T} \eta\, h_t(\mathbf{x}),$$

where η is the learning rate and h_t are the leaf-wise regression trees fitted via gradient boosting on the negative gradients of the loss function. LightGBM efficiently captures non-linear interactions among sparse EHR features and serves as a validated baseline for structured data.

This setup allows us to compare the EHR-AE approach with a literature-based baseline (EHR-GBM), demonstrating the benefits of learned dense embeddings maintaining at the same time methodological continuity with prior work.

Radiology reports For the Unimodal Radiology Report experiments, a novel hierarchical text encoder that leverages token-level embeddings $\mathbf{t}_{i,j}$ produced by Clinical-Longformer was developed. To capture both local and global semantics within each report, a two-level hierarchical attention mechanism was designed. At the first level, token embeddings $\mathbf{t}_{i,j}$ within each sentence i are aggregated according to learned attention weights, emphasizing clinically relevant terms and producing the sentence embedding $\mathbf{s}_i$:

$$\mathbf{s}_i = \sum_{j=0}^{L} w_{i,j} \mathbf{t}_{i,j}, \tag{5.1}$$

where $w_{i,j}$ is the weight learnt by the soft attention for the token embedding $\mathbf{t}_{i,j}$, L is the total number of tokens in the input sentence and i ranges from 1 to the total number of sentences

N in the radiology report. Since each sentence may contain a variable number of tokens and each radiology report consists of a different number of sentences, both L and N are defined dynamically at runtime for each mini-batch. Specifically, L is set to the maximum number of tokens in a single sentence across all sentences in the minibatch, whereas N is defined as the maximum number of sentences across all radiology reports in the mini-batch. To ensure uniform tensor shapes in the mini-batch, we use padding with all-zero vectors of size 768, which receive a weight of zero from both the token-level and sentence-level attention layers, ensuring they do not contribute to the final representation. It is worth nothing that $\mathbf{s}_i$ has the same size 768 of the token embedding.

At the second level, sentence representations were combined through the Sentence Attention layer, that computes:

$$\mathbf{x} = \sum_{i=0}^{N} w_i \mathbf{s}_i, \tag{5.2}$$

where w_i is the weight learnt by the soft attention for the sentence embedding $\mathbf{s}_i$, N is the total number of sentences in the radiology report. It provides a comprehensive representation of the radiology report. Importantly, this layer employs a soft attention mechanism that shares its weights with the token-level attention layer. This sharing is feasible because both token and sentence embeddings have the same dimensionality (768) making it possible to use a single soft attention module implemented as a fully connected layer of size 768. Weight sharing serves two main purposes: (i) it reduces the total number of trainable parameters, helping to prevent overfitting on limited data and lowering computational cost; and (ii) it improves feature transferability across layers, since the clinical relevance of a sentence is inherently related to the clinical relevance of its constituent entities. The final outcome is a weighted average vector, $\mathbf{x}$ of size 768. It acts as the feature vector for the MLP classification network, whose architecture is detailed in Table 5.3.

CTPA scans For the Unimodal Image (CTPA) model, ResNetV2-101 slices representations $\mathbf{c}_i$ of size 6145 were processed with a bidirectional GRU to capture sequential dependencies across slices. Then, a hybrid attention-and-max pooling mechanism was applied to aggregate the sequence into a single patient-level representation $\mathbf{u}$:

$$\mathbf{u} = \left[\max_{i=0,\ldots,R} \mathbf{c}_i \;\middle\|\; \sum_{i=0}^{R} w_i \mathbf{c}_i \right], \tag{5.3}$$

where $\max_{i=0,\ldots,R} \mathbf{c}_i$ denotes the *max pooling* operation over the resulting GRU slice-level embeddings, $\sum_{i=0}^{R} w_i \mathbf{c}_i$ corresponds to the attention-based aggregation (R is the total number of slices in the CTPA scan, w_i is a scalar weight reflecting the contribution of the i-th slice,

Layer	Output Size	Notes / Activation / Dropout
Input	`enc_size`	Embedding
Linear	512	Fully connected
ReLU	512	Activation
Dropout	512	$p = 0.2$
Linear	128	Fully connected
ReLU	128	Activation
Dropout	128	$p = 0.2$
Linear	1	Fully connected output layer (logit)

Table 5.3: Architecture of the MLP classification head.

and $\mathbf{c}_i$ represents the feature embedding of the i-th slice), whereas the operator $\|$ indicates the concatenation of the two resulting vectors. The unimodal CTPA model adopts the established architecture from INSPECT [63], providing a validated baseline for structured comparison with other modalities. The result is a CTPA representation $\mathbf{u}$ of dimension 12290, which is then fed into the MLP classification network, whose architecture is detailed in Table 5.3.

Multimodal Fusion Models

To integrate complementary information, a range of fusion strategies were explored, considering all pairwise modality combinations and the joint use of all three modalities.

Early Fusion The trimodal Early Fusion strategy is illustrated in Figure 5.2. In this approach, the modality-specific embeddings are directly concatenated into a single feature vector, which is then passed to a multi-layer perceptron (MLP) for classification (see Table 5.3 for architectural details).

Formally, let

$$\mathbf{z} \in \mathbb{R}^{d_{\text{ehr}}}, \quad \mathbf{x} \in \mathbb{R}^{d_{\text{rep}}}, \quad \mathbf{u} \in \mathbb{R}^{d_{\text{img}}},$$

denote the embeddings extracted from structured EHR data, clinical reports, and CTPA images, respectively, where d_{ehr}, d_{rep}, and d_{img} represent the dimensionalities of the three modality-specific embeddings. The embedding $\mathbf{z}$ is obtained by applying either Truncated SVD or an Autoencoder to the sparse matrix of count-based EHR features, resulting in a compact representation of the structured data. The embedding $\mathbf{x}$ is derived from radiology reports processed with Clinical Longformer, followed by a two-level Hierarchical Attention Mechanism (token-level and sentence-level attention). Finally, the embedding $\mathbf{u}$ is com-

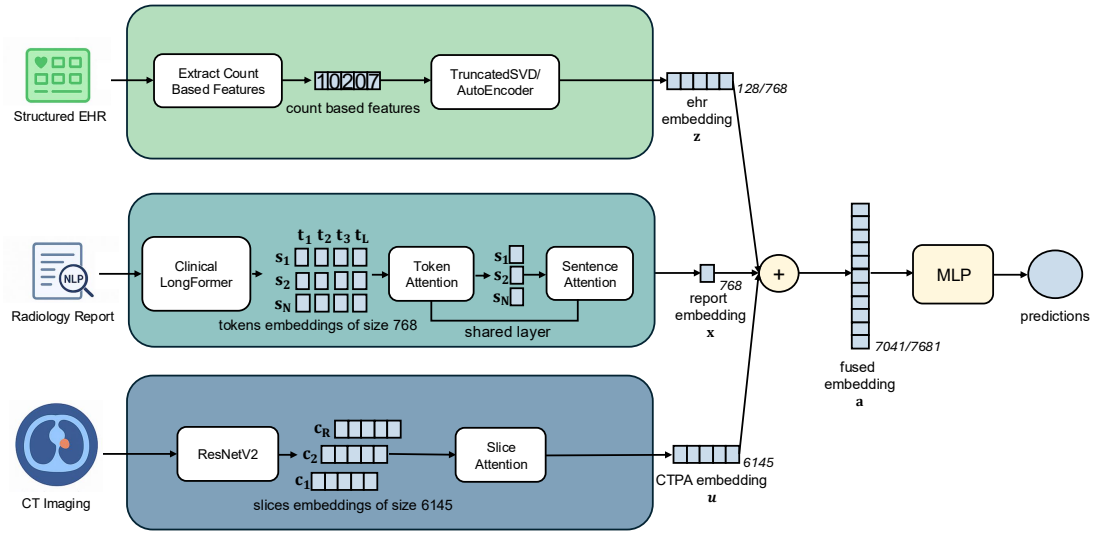

Figure 5.2: Overview of the *early fusion* method, where embeddings from EHR, clinical reports, and medical images are concatenated into a joint representation, subsequently processed by an MLP classifier.

puted from CT scans by first extracting slice-level representations using ResNetV2. These slice representations are then passed through an attention mechanism in which each slice is assigned a weight by a fully connected layer. The final CT embedding is given by the weighted aggregation:

$$\mathbf{u} = \sum_{i=1}^{R} w_i \mathbf{c}_i, \tag{5.4}$$

where w_i denotes the attention weight for slice embedding $\mathbf{c}_i$, and R is the total number of slices in the CTPA study. In contrast to the unimodal CTPA model, max pooling is omitted in the multimodal setting. This choice is motivated by the fact that (i) the attention mechanism already captures the relative importance of each slice, and (ii) removing max pooling reduces the dimensionality of the resulting embedding, limiting the number of trainable parameters in the fusion model.

The Early Fusion representation is then obtained by concatenation:

$$\mathbf{a} = [\,\mathbf{z} \,\|\, \mathbf{x} \,\|\, \mathbf{u}\,] \in \mathbb{R}^{d_{\text{ehr}}+d_{\text{rep}}+d_{\text{img}}},$$

where $\|$ denotes the concatenation operator. The fused vector is then passed to the MLP classifier:

$$\hat{\mathbf{y}} = \sigma(f_{\text{MLP}}(\mathbf{a}))\,,$$

Method	**Modalities**				**Embedding**
	EHR-TSVD	**EHR-AE**	**CT**	**Report**	
Unimodal	✓	-	-	-	128
Unimodal	-	✓	-	-	768
Unimodal	-	-	✓	-	12,290
Unimodal	-	-	-	✓	768
Bimodal	✓	-	✓	-	6,273
Bimodal	✓	-	-	✓	896
Bimodal	-	✓	✓	-	6,913
Bimodal	-	✓	-	✓	1,536
Bimodal	-	-	✓	✓	6,913
Trimodal	✓	-	✓	✓	7,041
Trimodal	-	✓	✓	✓	7,681

Table 5.4: Overview of fusion strategies used in the experiments. The table reports the included modalities (EHR, CT, Report) and the resulting embedding dimension for the early fusion strategy.

with $f_{\text{MLP}}(\cdot)$ denoting the feed-forward classification network and $\sigma(\cdot)$ the final sigmoid activation.

The dimensionalities of all modality-specific embeddings used in the Early Fusion strategy are reported in Table 5.4. In addition to the trimodal configuration, all bimodal combinations were also evaluated.

Intermediate Fusion This strategy leverages joint training with modality-specific encoders, whose latent representations are integrated through a cross-attention mechanism. Unlike Early Fusion, where embeddings are directly concatenated, here the model explicitly learns inter-modality interactions by aligning tokens across different representation spaces.

For EHR data, we use embeddings derived from the sparse matrix of count-based features, obtained either via Truncated SVD or the Autoencoder. For Radiology reports, we adopt embeddings produced by the two-level hierarchical attention mechanism ([118]), built upon representations generated by the Clinical LongFormer. Finally, for CT data, we utilize slice-level representations extracted using ResNetV2.

To ensure dimensional consistency across inputs, each modality was projected through an embedding layer that maps representations to a common size of 512. The following subsections distinguish between trimodal and bimodal fusion configurations.

For the trimodal setting, two intermediate fusion strategies were employed: Joint Fusion

with contrastive alignment (ARMOUR) and Cascaded Fusion (CROSS). These methods were selected for their complementary capabilities in modeling cross-modal interactions and for their suitability for multimodal clinical data.

In the first strategy, the ARMOUR framework proposed by [97] was adopted. ARMOUR enables robust bimodal fusion for clinical prediction tasks while preserving performance in scenarios with missing modalities. In the present configuration, ARMOUR was used to fuse CTPA imaging and radiology report embeddings, producing a joint image–text representation. This fused representation was subsequently combined with the EHR embedding, resulting in a trimodal fusion architecture.

To enable cross-modal interaction, each modality is first projected into a shared embedding space ("Embedding" in Figure 5.3) of equal dimensionality through modality-specific projection layers. Subsequently, modality-specific class tokens are introduced for each modality (e.g., imaging and reports) to act as compact global representations that summarize modality-level semantics. These class tokens are then refined via cross-attention layers, where one modality attends to the other (e.g., text attending to image features and vice versa), allowing the model to capture fine-grained interactions and contextual dependencies across modalities.

A contrastive alignment strategy, inspired by Momentum Contrast (MoCo, [53, 26]), is applied to these class tokens. The contrastive loss encourages paired embeddings from the same sample to be close in the latent space while pushing apart embeddings from different samples. This promotes semantic coherence between modalities and enhances robustness under missing-modality conditions, as the network can leverage complementary information from the available modalities.

After the cross-attention layers, modality-specific class tokens are fused through concatenation. In our adaptation, the resulting fused representation from the ARMOUR cross-attention and contrastive alignment stage is then concatenated with the EHR embedding. The unified multimodal vector is finally passed to the classifier for prognostic prediction (see Figure 5.3).

In the CROSS strategy, multimodal interactions are modelled through a hierarchical sequence of cross-attention operations rather than a joint or parallel fusion. Formally, the cross-attention operation is defined as:

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\mathbf{d}_k}}\right)\mathbf{V},$$

where $\mathbf{Q} = \mathbf{X}_Q\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}_K\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}_V\mathbf{W}_V$, with $\mathbf{X}_Q, \mathbf{X}_K, \mathbf{X}_V$ denoting the query, key, and value sequences from different modalities.

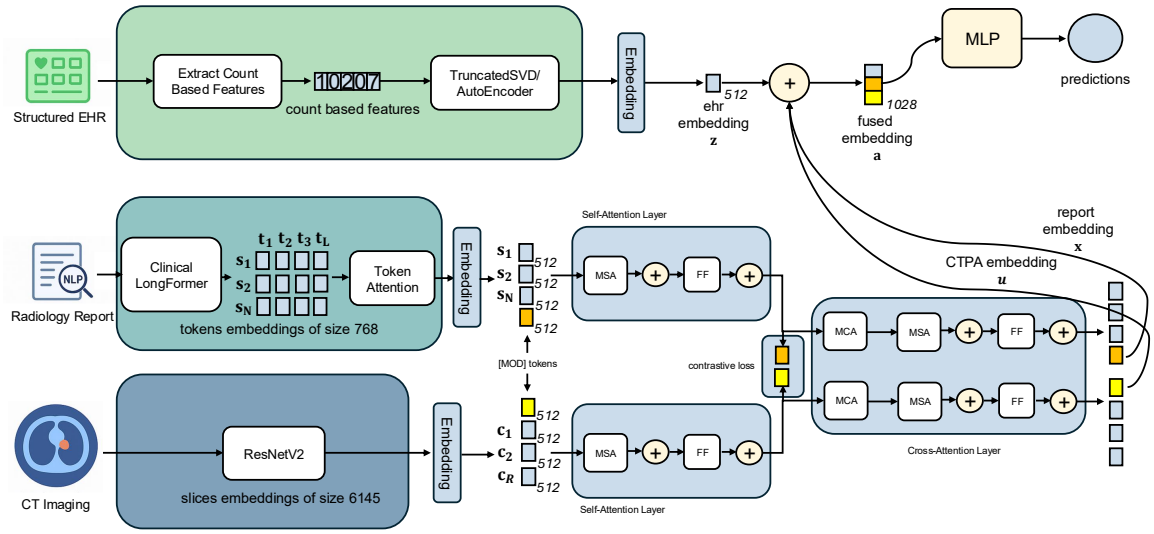

Figure 5.3: Conceptual illustration of the ARMOUR workflow, showing cross-attention blocks between modality-specific encoders, the contrastive alignment mechanism, and the final fusion with the EHR embedding.

To capture richer interactions across multiple subspaces, this operation was extended to a multi-head cross-attention (MHCA) mechanism:

$$\begin{aligned} \mathrm{MHCA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathrm{Concat}(\mathbf{h}_1, \ldots, \mathbf{h}_H)\, \mathbf{W}^O, \\ \mathbf{h}_i &= \mathrm{CrossAttn}(\mathbf{Q}\mathbf{W}_i^Q,\, \mathbf{K}\mathbf{W}_i^K,\, \mathbf{V}\mathbf{W}_i^V), \end{aligned}$$

where $\mathbf{H}$ denotes the number of heads.

In this approach, the EHR embedding, represented as a single query token, acts as the central information carrier that sequentially integrates complementary information from other modalities. Specifically, in the first stage, the EHR token attends to one modality (e.g., imaging), acquiring modality-specific contextual features and generating an enriched representation. This updated token then serves as the query in a second cross-attention step, attending to another modality (e.g., textual reports), thereby incorporating additional semantic cues.

This cascaded design allows the model to progressively refine the EHR representation through modality-specific conditioning, enabling an interpretable flow of information across modalities and reducing the complexity of joint optimization. Unlike ARMOUR, which align modalities in a shared latent space simultaneously, CROSS enforces a sequential dependency that mimics a stepwise reasoning process, first grounding the representation in one modality before enriching it with another. The final token generated by the cascaded process thus

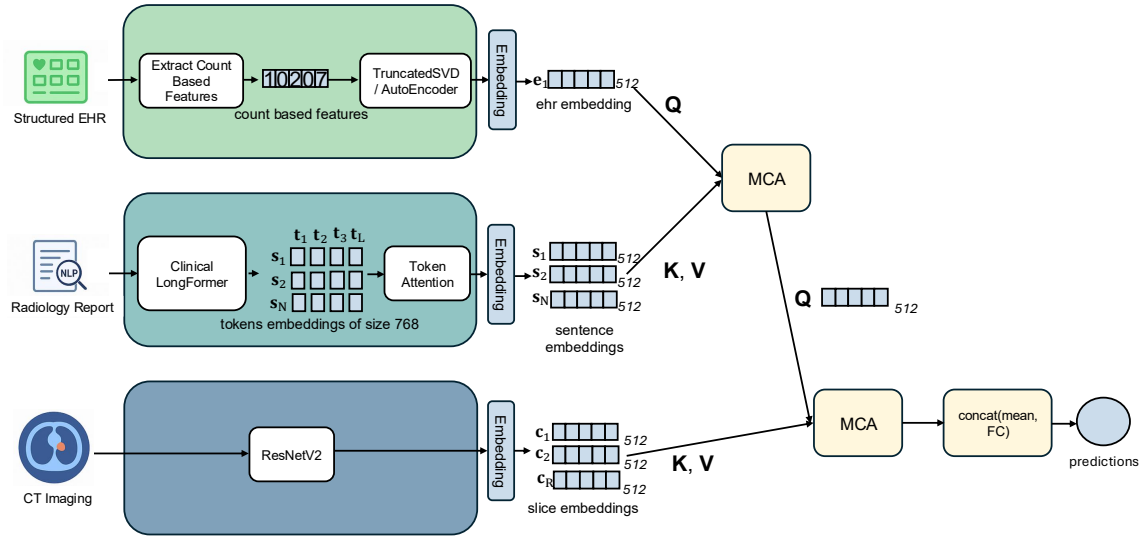

Figure 5.4: Illustration of a case of trimodal *intermediate fusion* approach: A cascaded fusion mechanism is employed, where the EHR embedding sequentially attends to the other modalities through a Multi-Head Classification (MCA) layer prior to the final classification.

encodes aggregated multimodal information and is subsequently passed to the classification layer for prognostic prediction, as illustrated in Figure 5.4.

Alternative cascade configurations were also explored, varying the ordering of modalities within the two-stage attention process. In the first configuration, the EHR representation acted as the query and sequentially attended to the report and then to the image, producing a single enriched token that was directly passed to the classifier. In the second configuration, the cascade began with image tokens serving as the query attending to the report, followed by attention to the EHR representation. In this setup, the full token sequences were preserved rather than collapsed into a single token, and the resulting sequence was averaged to obtain a mean embedding for classification.

These alternative designs made it possible to assess how different query–key/value hierarchies influence the propagation of cross-modal information and, ultimately, impact predictive performance.

For bimodal setting, experiments were carried out using only the CROSS scheme with a single Multi-head Cross-Attention (MCA) layer, where the query sequence was taken from one modality and the key–value pairs from the other. The evaluated modality pairs correspond to those considered in the early fusion setting (Table 5.4); however, note that the resulting embeddings in this CROSS scheme have different dimensionalities from those reported for early fusion. In the specific case of pairs with EHR data, which is represented by a single token, this token was directly used as the query. Consequently, the cross-attention

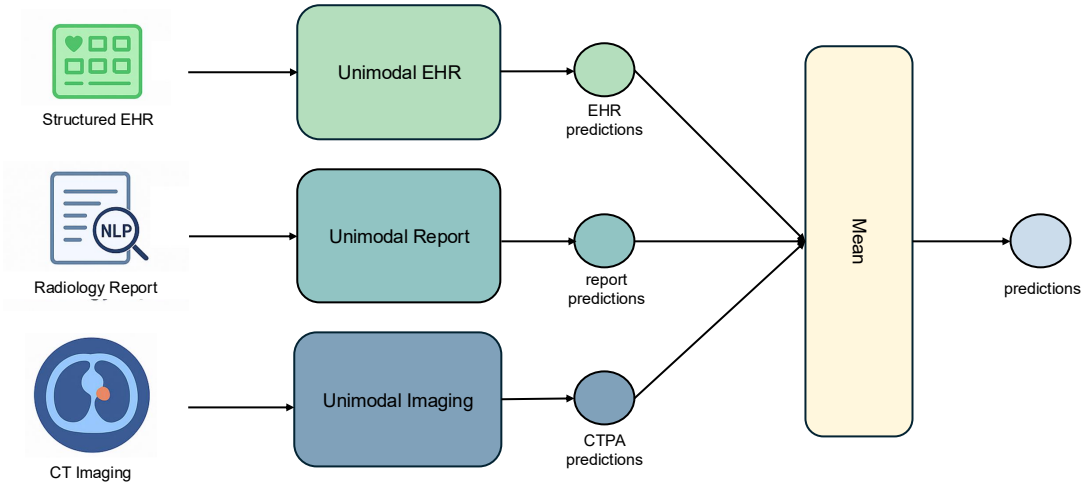

Figure 5.5: Illustration of the *late fusion* method, where predictions from unimodal models are averaged to obtain the final decision.

operation yields an output sequence reduced to a single token, which is subsequently passed to the classification layer. In all other cases, the output sequence produced by the cross-attention mechanism was averaged across tokens, and the resulting mean representation was fed into the classification layer.

Late Fusion Ensemble averaging of unimodal predictions was performed, as illustrated in Figure 5.5, by combining the outputs of the three unimodal models (EHR, report, and imaging) through the averaging of their predicted probabilities. The final predicted class is then obtained by applying a threshold of 0.5 for a binary outcome. Both bimodal and trimodal combinations were considered. For the bimodal case, was used the same pairs as in the Early Fusion reported in Table 5.4; however, note that the resulting embeddings in this Late fusion scheme have different dimensionalities from those reported for early fusion.

5.3 Experimental Setup

Unimodal and multimodal models were trained and evaluated using the patient-level hold-out splits defined in [63], allocating 80% of the data to the training set, 5% to the validation set, and 15% to the test set. Our effective cohort comprises a subset of the original INSPECT dataset, totaling 23248 CTPA studies. Details on the number of samples selected for each prognostic horizon are provided in Section 5.2.1. Samples lacking a valid reference time point (e.g., CT scan timestamp) were excluded to ensure that only pre-exam EHR events are used for prognostic modeling.

All experiments were implemented in `PyTorch` and executed on an NVIDIA A40 GPU with 48 GB of VRAM. Models were trained for up to 50 epochs using the AdamW optimizer with a weight decay of 0.02, a learning rate of 0.001, and a batch size of 128. The cross-entropy loss was employed as the training objective. To better characterize performance variability, each experiment was repeated over five independent runs with different random seeds.

For the EHR-GBM model, rather than relying on the default configuration, we performed a targeted hyperparameter search over a compact grid including max depth [3, 6, –1], learning rate [0.02, 0.1, 0.5], and number of leaves [10, 25, 100].

Model performance was evaluated using the Matthews Correlation Coefficient (MCC), the Area Under the Receiver Operating Characteristic curve (AUROC), and the Area Under the Precision-Recall curve (AUPRC). The MCC is defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC provides a balanced measure of prediction quality even in the presence of class imbalance, as it considers all four entries of the confusion matrix.

These metrics were chosen because the dataset is inherently imbalanced: mortality events are much less frequent than survival outcomes, especially at shorter prognostic horizons. On the one hand, accuracy can be misleading in such scenarios, as a model could achieve high accuracy by simply predicting the majority class. Similarly, F1-score does not account for true negatives, even though it captures the harmonic mean of precision and recall. On the other hand, MCC incorporates all entries of the confusion matrix, AUROC provides a threshold-independent measure of discriminative ability, and AUPRC emphasizes the model's performance on the minority positive class, which is critical for prognostic prediction.

5.4 Results and Discussion

This section presents the results of unimodal and multimodal experiments across three prognostic time horizons (1-, 6-, and 12-month mortality). Given the pronounced class imbalance in the INSPECT dataset (Table 1), with positive mortality rates of 5.45%, 11.54%, and 14.59% for 1-, 6-, and 12-month prediction tasks, respectively, the MCC was prioritized as the primary evaluation metric for model comparison and discussion, whereas additional metrics (e.g., AUROC, F1-score) are reported in the Appendix for completeness. MCC effectively balances true positives, true negatives, false positives, and false negatives, making

it ideal for assessing prognostic models in imbalanced settings like PE.

The findings of this thesis are structured around five key research questions, which collectively clarify the individual contribution of each modality, the behavior of different fusion strategies, and the broader implications for multimodal prognostic modeling in PE:

1. How well do unimodal models perform individually?
2. Does multimodal fusion improve predictive performance over the strongest unimodal baseline?
3. Which fusion strategy is most robust across the three prognostic horizons?
4. What are the potential pitfalls of complex fusion mechanisms such as CROSS and ARMOUR?
5. What potential clinical insights can multimodality provide?

This section address these questions sequentially, beginning with unimodal performance and progressively incorporating more complex multimodal interactions, to uncover how each data source and fusion mechanism influences predictive accuracy and robustness.

Table 5.5: MCC of unimodal models across 1-, 6-, and 12-month mortality prediction tasks. Results are reported as mean ± standard deviation over five runs. Best MCC values for each time horizon are highlighted in **bold**

Model	1-Month	6-Month	12-Month
EHR-GBM	0.258 ± 0.059	**0.367 ± 0.072**	**0.454 ± 0.023**
EHR-AE	**0.269 ± 0.013**	0.352 ± 0.016	0.395 ± 0.008
Report	0.242 ± 0.008	0.329 ± 0.030	0.354 ± 0.014
Image	0.221 ± 0.010	0.297 ± 0.016	0.304 ± 0.028

(1) How well do unimodal models perform individually? To establish baseline performance, unimodal models were evaluated separately on EHR, report, and imaging data across the three prognostic horizons (1-, 6-, and 12-month mortality). Table 5.5 reports the results in terms of MCC, AUC, and AUPRC.

For structured EHR data, two models were assessed: a Gradient Boosting Machine (GBM) and a Deep Autoencoder (AE). The EHR–AE achieved the highest mean MCC at the 1-month horizon (0.269), whereas EHR–GBM obtained the highest mean MCC at

both the 6-month (0.367) and 12-month (0.454) horizons. In addition, EHR–GBM consistently exhibited strong AUC and AUPRC values across all tasks, underscoring the predictive value of structured clinical variables.

The textual report model (Report in Table 5.5), based on Clinical-Longformer embeddings and hierarchical attention, achieved moderate performance, with mean MCC values of 0.242, 0.329, and 0.354 at 1-, 6-, and 12-month horizons, respectively. Meanwhile, the imaging-only model (Image in Table 5.5), relying on ResNetV2 features extracted from CT volumes, exhibited the lowest performance across all horizons, with mean MCC scores of 0.221, 0.297, and 0.304.

To assess whether the observed differences in MCC across the four unimodal baselines were statistically meaningful, paired one-tailed t-tests over the five independent runs were conducted for each prognostic horizon and applied Benjamini–Hochberg correction to control the false discovery rate ([13]). This analysis revealed three distinct patterns depending on the prediction window.

At the 1-month horizon, the EHR-AE clearly emerged as the strongest performer, significantly outperforming both the report-only model ($t = 3.96$, FDR-adjusted $p = 0.011$) and the imaging-only baseline ($t = 6.54$, FDR-adjusted $p = 0.0008$). Its performance, however, remained statistically indistinguishable from that of EHR–GBM ($p = 0.81$), suggesting that the two EHR-based approaches offer comparable predictive power in the short term.

At the 6-month horizon, the landscape became more balanced. No significant differences emerged between either of the EHR-based models and the report-only system. Nonetheless, the EHR–AE maintained a clear advantage over imaging ($t = 5.44$, FDR-adjusted $p = 0.004$), while the marginal superiority of EHR–GBM over imaging did not survive FDR correction ($p = 0.32$).

At the 12-month horizon EHR–GBM established itself as the dominant unimodal approach . It significantly outperformed the Autoencoder ($t = 5.42$, FDR-adjusted $p = 0.0013$), the report-only model ($t = 8.30$, FDR-adjusted $p = 0.0002$), and the imaging model ($t = 9.26$, FDR-adjusted $p = 0.0001$). Even so, the EHR-AE continued to retain a significant edge over both the report-only ($t = 5.69$, FDR-adjusted $p = 0.002$) and image-only ($t = 6.99$, FDR-adjusted $p = 0.0008$) baselines, confirming the consistent strength of structured EHR information across all horizons.

Overall, the findings delineate a clear and coherent trend: structured EHR data offer the strongest and most reliable unimodal signal for mortality prediction. The EHR-AE proves most effective in the short term, whereas the EHR–GBM progressively asserts itself as the leading baseline at longer horizons. Nevertheless, the steady, though comparatively weaker, predictive contribution of textual and imaging modalities underscores their value

as complementary sources of information, supporting their integration within multimodal fusion frameworks.

(2) Does multimodal fusion improve predictive performance over the strongest unimodal baseline? To evaluate the benefit of multimodal integration, each fusion strategy was compared against the strongest unimodal baselines, the EHR-AE at the 1-month horizon (mean MCC of 0.269) and the EHR–GBM at the 6- and 12-month horizons (mean MCCs of 0.367 and 0.454, respectively). Table A.2 reports MCC, AUC, and AUPRC for all unimodal models and the early, late, CROSS, and ARMOUR fusion schemes.

Late fusion strategies that incorporated the EHR–GBM model predictions consistently delivered the strongest results. Across the three prognostic horizons, different combinations of modalities emerged as the most effective. At the 1-month horizon, the best-performing configuration was the late fusion of clinical reports with EHR–GBM, reaching an MCC of 0.399 ± 0.050, which corresponds to a notable gain of $+0.130$ over the EHR-AE. Moving to the 6-month horizon, performance peaked with a fusion model that integrated reports, imaging, and EHR–GBM, achieving an MCC of 0.479 ± 0.011 and improving upon the EHR–GBM baseline by $+0.112$. At the 12-month horizon, the most effective strategy combined imaging with EHR–GBM, yielding an MCC of 0.497 ± 0.011, a $+0.043$ gain over the GBM alone.

Notably, the three-modality late fusion model (Reports, Image, EHR–GBM) maintained strong performance across all horizons, with MCC values of 0.362, 0.479, and 0.488, reinforcing the value of integrating complementary information sources in multimodal frameworks.

Early fusion yielded more modest gains. The best early fusion variants (EHR–TSVD + Report + Image and EHR–TSVD + Report) reached MCC values up to 0.426 at 12-months but never significantly surpassed EHR–GBM on any horizon. Most CROSS-fusion and ARMOUR-fusion approaches performed poorly, with many MCC scores below or close to the unimodal report or image baselines, showing that sequential prediction and complex attention routing may introduce noise and error propagation in this setting.

To assess whether multimodal fusion genuinely contributed additional predictive value, we focused on all late fusion strategies that incorporated EHR–GBM predictions and evaluated them against the strongest EHR-only baseline. For each horizon, we applied paired one-tailed t-tests on MCC across the five independent runs, followed by FDR correction using the Benjamini–Hochberg procedure ([13]). This analysis, summarized in Table 5.7, revealed a marked contrast with respect to other integration approaches: late fusion configurations involving EHR–GBM displayed a consistent and statistically robust advantage, signalling a clear shift in performance.

Table 5.6: MCC of unimodal and multimodal models across 1-, 6-, and 12-month mortality prediction tasks (mean ± std over five runs). Best values in **bold**. In Cross Fusion models, "→" indicates the order of modality integration.

Model	1-Month	6-Month	12-Month
Unimodal			
EHR-GBM	0.258 ± 0.059	0.367 ± 0.072	0.454 ± 0.023
EHR-AE	0.269 ± 0.013	0.352 ± 0.016	0.395 ± 0.008
Report	0.242 ± 0.008	0.329 ± 0.030	0.354 ± 0.014
Image	0.221 ± 0.010	0.297 ± 0.016	0.304 ± 0.028
Early Fusion			
EHR-AE, Report	0.296 ± 0.014	0.371 ± 0.018	0.407 ± 0.009
EHR-AE, Image	0.302 ± 0.011	0.380 ± 0.016	0.405 ± 0.010
EHR-TSVD, Report	0.280 ± 0.027	0.392 ± 0.006	0.419 ± 0.010
EHR-TSVD, Image	0.274 ± 0.013	0.376 ± 0.011	0.390 ± 0.017
Report, Image	0.244 ± 0.024	0.350 ± 0.016	0.366 ± 0.012
EHR-AE, Report, Image	0.302 ± 0.008	0.378 ± 0.014	0.398 ± 0.007
EHR-TSVD, Report, Image	0.293 ± 0.009	0.393 ± 0.008	0.426 ± 0.008
Late Fusion			
Reports, EHR-AE	0.286 ± 0.015	0.388 ± 0.019	0.424 ± 0.005
Reports, EHR-GBM	**0.399 ± 0.050**	0.472 ± 0.011	0.494 ± 0.008
Image, EHR-AE	0.297 ± 0.017	0.395 ± 0.016	0.426 ± 0.004
Image, EHR-GBM	0.374 ± 0.025	0.467 ± 0.031	**0.497 ± 0.011**
Reports, Image	0.275 ± 0.015	0.375 ± 0.011	0.398 ± 0.012
Reports, Image, EHR-AE	0.331 ± 0.014	0.409 ± 0.014	0.444 ± 0.008
Reports, Image, EHR-GBM	0.362 ± 0.016	**0.479 ± 0.011**	0.488 ± 0.002
Cross Fusion			
EHR-TSVD → Image → Report	0.240 ± 0.023	0.321 ± 0.017	0.351 ± 0.028
EHR-AE → Image → Report	0.227 ± 0.011	0.327 ± 0.019	0.404 ± 0.022
EHR-TSVD → Report	0.247 ± 0.021	0.350 ± 0.028	0.377 ± 0.035
EHR-TSVD → Image	0.181 ± 0.024	0.294 ± 0.063	0.294 ± 0.119
EHR-AE → Report	0.285 ± 0.032	0.363 ± 0.019	0.391 ± 0.039
Image → Report	0.236 ± 0.025	0.341 ± 0.016	0.361 ± 0.013
EHR-TSVD → Report → Image	0.084 ± 0.032	0.146 ± 0.052	0.087 ± 0.051
EHR-AE → Report → Image	0.124 ± 0.083	0.204 ± 0.158	0.165 ± 0.184
Image → Report → EHR-AE	0.305 ± 0.007	0.372 ± 0.023	0.397 ± 0.010
Image → Report → EHR-TSVD	0.250 ± 0.021	0.346 ± 0.017	0.388 ± 0.013
Armour Fusion			
EHR-TSVD, Image, Report	0.239 ± 0.027	0.360 ± 0.030	0.391 ± 0.024
EHR-AE, Image, Report	0.284 ± 0.023	0.372 ± 0.015	0.420 ± 0.012

Table 5.7: Statistical comparison of late fusion strategies incorporating EHR–GBM against the EHR-only baseline. Paired one-tailed t-tests were applied on MCC across five independent runs, with FDR correction using the Benjamini–Hochberg procedure.

Fusion Strategy	**1-month**	**6-month**	**12-month**
Reports + EHR–GBM	$t = 4.08$, FDR-$p = 0.0020$	$t = 3.22$, FDR-$p = 0.0066$	$t = 3.67$, FDR-$p = 0.0034$
Image + EHR–GBM	$t = 4.05$, FDR-$p = 0.0021$	$t = 2.85$, FDR-$p = 0.0116$	$t = 3.77$, FDR-$p = 0.0029$
Reports + Image + EHR–GBM	$t = 3.80$, FDR-$p = 0.0029$	$t = 3.44$, FDR-$p = 0.0048$	$t = 3.29$, FDR-$p = 0.0059$

In conclusion, simple decision-level late fusion of independent modality-specific predictions with the strong EHR–GBM model predictions consistently and significantly improves mortality prediction across all three prognostic horizons. The combination of structured EHR data with either narrative radiology reports, imaging features, or both yields the highest absolute performance, with gains of up to +14.1% MCC at 1-month and +11.2% MCC at 6-months. These results point out that radiology reports and CT imaging encode complementary prognostic information that is most effectively exploited through late fusion rather than joint representation learning or complex sequential architectures.

(3) Which fusion strategy is most robust across the three prognostic horizons? To evaluate the robustness of each fusion strategy across the 1-, 6-, and 12-month mortality prediction tasks, mean MCC performance, ranking consistency, and statistical superiority over competing strategies were considered.

To provide a rigorous statistical synthesis of fusion strategy performance, Table 5.8 reports pairwise win rates across all models, metrics, and horizons. Late trimodal fusion (LTri.) dominates, achieving near-perfect win rates (0.8–1.0) with 100% significance against most categories and 86% significance vs. unimodal models. Late bimodal (LBi.) also excels, particularly against early and CROSS fusion.

Early fusion shows moderate strength against unimodal models (0.7 wins, 85-92% significance) but is consistently outperformed by late fusion. ARMOUR trimodal wins against CROSS and unimodal models but never beats late fusion (0.0 wins). Cross fusion (CTri.) performs the weakest among the evaluated strategies, exhibiting win rates at or below 0.3 and generally lacking statistical significance in most comparisons.

These results statistically confirm that late fusion, especially trimodal, is the most robust and generalizable strategy, even when accounting for metric-specific fluctuations (e.g., occasional bimodal MCC wins at 1- and 12-month). Its superiority can be attributed to its modular design: each modality is modeled independently at its optimal capacity, particularly the strong EHR–GBM component, and only calibrated predictions are combined at

Fusion Type	**Uni.**	**EBi.**	**ETri.**	**LBi.**	**LTri.**	**CBi.**	**CTri.**	**ATri.**
Uni.	–	0.3 (60%)	0.3 (100%)	0.2 (100%)	0.2 (73%)	0.6 (11%)	0.8 (66%)	0.6 (66%)
EBi.	0.7 (85%)	–	0.3 (63%)	0.1 (0%)	0.0 (0%)	0.7 (61%)	0.9 (77%)	0.8 (77%)
ETri.	0.7 (92%)	0.7 (44%)	–	0.2 (11%)	0.0 (0%)	0.8 (81%)	0.9 (88%)	0.9 (58%)
LBi.	0.8 (100%)	0.9 (88%)	0.8 (72%)	–	0.2 (78%)	1.0 (100%)	1.0 (100%)	1.0 (68%)
LTri.	0.8 (100%)	1.0 (100%)	1.0 (100%)	0.8 (79%)	–	1.0 (100%)	1.0 (100%)	1.0 (100%)
CBi.	0.4 (77%)	0.2 (49%)	0.2 (85%)	0.0 (0%)	0.0 (0%)	–	0.8 (78%)	0.5 (61%)
CTri.	0.2 (63%)	0.1 (83%)	0.1 (67%)	0.0 (0%)	0.0 (0%)	0.2 (27%)	–	0.3 (47%)
ATri.	0.4 (90%)	0.2 (71%)	0.1 (67%)	0.0 (50%)	0.0 (0%)	0.5 (70%)	0.7 (74%)	–

Table 5.8: Pairwise win comparisons across fusion strategies, aggregated over all time horizons and metrics (AUC, MCC, AUPRC). Each cell reports the proportion of wins by the row model vs. the column model, with the percentage of statistically significant wins (FDR-adjusted $p < 0.005$) in parentheses. **Abbreviations:** Uni. = Unimodal; EBi. = Early Bimodal; ETri. = Early Trimodal; LBi. = Late Bimodal; LTri. = Late Trimodal; CBi. = Cross Bimodal; CTri. = Cross Trimodal; ATri. = Armour Trimodal.

the decision level. This minimizes interference between heterogeneous data types, reduces overfitting, and preserves the predictive power of structured EHR data while extracting complementary information from reports and imaging.

Across the three prediction horizons, the optimal late fusion configurations are: Radiology Reports with EHR–GBM for 1-month mortality, Radiology Reports, CT Imaging, and EHR–GBM for 6-month mortality, and CT Imaging with EHR–GBM, or alternatively all three modalities, for 12-month mortality, yielding very similar performance. These simple, interpretable ensembles consistently achieve the highest predictive performance with statistically significant improvements over other fusion strategies.

(4) What are the pitfalls of complex fusion mechanisms (CROSS and ARMOUR)? Although advanced fusion strategies such as CROSS Fusion and ARMOUR Fusion aim to capture intricate inter-modal interactions through attention-based or bidirectional mecha-

nisms, our results reveal significant pitfalls that undermine their effectiveness and robustness in clinical mortality prediction.

Cross Fusion exhibits severe performance degradation and instability, particularly in multi-step cascading configurations. Despite using the same modalities, performance varies dramatically with fusion order. For instance, the sequence EHR TSVD $\rightarrow$ Report $\rightarrow$ Image yields MCC values of only 0.084 ± 0.032 (1-month), 0.146 ± 0.052 (6-month), and 0.087 ± 0.051 (12-month), drastically worse than unimodal baselines. This sensitivity to ordering indicates that the attention mechanism may amplify noise or introduce destructive interference when weaker modalities (e.g., Image) are used as queries or keys in early stages. Furthermore, high variance across runs (e.g., ± 0.184 in EHR Autoencoder $\rightarrow$ Report $\rightarrow$ Image at 12-months) suggests poor generalization and overfitting to spurious correlations in training splits.

Similarly, ARMOUR Fusion, designed for bidirectional modality refinement, fails to outperform simpler strategies. Even in its full three-modality form (EHR, Image, Report), it achieves only moderate MCCs (0.239–0.284 at 1-month, up to 0.420 at 12-months) and never surpasses Late Fusion. The incremental refinement steps appear to dilute discriminative signals from the dominant EHR-GBM predictor, adding computational overhead without proportional gain.

A key shared pitfall is the risk of over-parameterization in high-dimensional heterogeneous clinical data. Both CROSS and ARMOUR introduce numerous attention heads and fusion layers, which, although expressive, increase the risk of overfitting, especially given the limited size of medical cohorts and class imbalance in mortality tasks. In contrast, Late Fusion preserves modality-specific optimization and combines predictions at the decision level, avoiding interference.

Moreover, the behavior of the more complex fusion mechanisms reveals a lack of robustness across prediction horizons. Whereas Late Fusion improves steadily from 1 to 12 months, both CROSS and ARMOUR exhibit irregular and often unstable performance, with several configurations deteriorating sharply at longer horizons despite the availability of richer longitudinal information. This pattern suggests that these intricate fusion schemes may struggle to scale with increasing predictive difficulty.

In practical terms, these shortcomings make CROSS and ARMOUR unreliable candidates for real-world deployment. Late Fusion, by contrast, stands out as the most effective strategy: it is simple, stable across horizons, and aligns naturally with clinical workflows, confirming its role as the preferred approach.

(5) What potential clinical insights can multimodality provide? Multimodal models integrating EHR, radiology reports, and CT images could offer rich clinical insights beyond unimodal predictions, revealing complementary risk signals and temporal patterns in mortality prediction. These insights may support personalized clinical decision-making.

First, EHR dominates unimodal performance across all horizons (MCC: 0.269 at 1-month, 0.367 at 6-month, 0.454 at 12-months), confirming that structured clinical variables, such as vital signs, lab results, and comorbidities, carry the strongest and most stable prognostic signal. This underscores the foundational role of electronic health records in risk stratification.

However, multimodal Late Fusion consistently outperforms EHR alone, indicating additive predictive value from unstructured modalities.

At the 1-month horizon, combining Reports with EHR-GBM yields a sizeable gain of +0.130 MCC (from 0.269 to 0.399). This improvement indicates that radiology reports may encode acute, clinically actionable findings, such as signs of respiratory distress or emergent complications, that are only partially reflected in structured fields.

Over longer horizons, the contribution of imaging becomes more prominent. At 6 and 12 months, the addition of CT features within the Reports + Image + EHR-GBM ensemble produces further gains (+0.112 to +0.034 MCC). This suggests that CT scans may capture slowly evolving anatomical patterns, chronic vascular changes, cardiopulmonary remodeling, and other structural markers, that carry long-term prognostic value.

Taken together, these results reveal a clear temporal structure in multimodal prediction: short-term mortality is best informed by acute descriptors encoded in narrative reports, whilst long-term risk increasingly depends on structural information visible in imaging. The complementary nature of these modalities underscores the value of multimodal fusion for a holistic, time-sensitive assessment of patient trajectory.

Furthermore, the failure of complex fusion (CROSS/ARMOUR) despite access to the same data highlights that not all inter-modal interactions are beneficial. Destructive fusion in cascading attention (e.g., EHR $\rightarrow$ Image $\rightarrow$ Report) may propagate imaging noise into textual reasoning, reducing reliability. In this context, clinical multimodal systems may benefit from modular, decision-level integration (Late Fusion) to preserve signal fidelity, though this may not generalize to all tasks or datasets.

Finally, in the best-performing models, AUPRC increases from 0.394 at 1 month to 0.581 at 12 months (see A.2 for full results), suggesting that class separability becomes more pronounced for longer prediction horizons. Multimodal architectures seem to exploit this widening gap particularly well: they leverage early warning cues embedded in radiology reports while integrating anatomical information that becomes apparent in the CT scans at

the time of the visit.

This evolution across time also suggests how such systems could be deployed in clinical practice. In the immediate, 1-month window, radiology reports combined with EHR data may offer rapid and actionable insight, supporting early risk stratification in acute scenarios. As the horizon extends to 6- and 12-months, imaging features may gain increasing predictive relevance, making them valuable for monitoring disease progression and gauging treatment response. From an implementation standpoint, late-fusion ensembles are especially practical, as they combine modality-specific predictions while remaining robust to variability in input data availability.

In essence, multimodal models reveal how different data sources contribute complementary prognostic value at different stages of the patient trajectory, turning heterogeneous clinical information into coherent, horizon-specific decision support.

Comparison with the Original INSPECT Study A direct, point-by-point comparison with [63] is not possible, as several methodological and practical differences separate our work from the original study. Three elements, in particular, shape this divergence.

First, our analysis is conducted on a reduced version of the INSPECT cohort. Patients lacking a reliable reference timestamp (CT acquisition time) had to be removed, since structured EHR features and prognostic labels could not be aligned consistently. This necessary filtering both reduced the available data and slightly altered class balance.

Second, although the MOTOR model ([141]), a time-to-event foundation model designed by the creators of the INSPECT dataset to extract temporal representations from structured EHR data, is officially released, its use remains out of reach in practice. The absence of operational code, documentation, and working preprocessing pipelines prevents us from reproducing its reported performance (AUC ≈ 0.924) or from incorporating it as a comparative baseline.

Third, differences in evaluation metrics complicate direct comparisons. [63] reported only AUC, which can be sensitive to class imbalance and potentially misleading in rare-event settings ([132]). In our experiments, for example, the EHR–GBM model, the same EHR-based Gradient Boosting Machine baseline used in the INSPECT study, achieves a higher AUC at 1 month (0.926) than our best late-fusion model (0.884), yet its MCC is substantially lower (0.258 vs. 0.399). This illustrates how AUC and MCC can provide complementary perspectives, and why MCC is particularly informative in the presence of class imbalance.

Despite these constraints, we were able to reproduce the main INSPECT baselines under identical preprocessing conditions. For the unimodal CT model, our AUCs reached 0.800, 0.790, and 0.781 at the 1-, 6-, and 12-month horizons, closely matching the original results

(0.794, 0.755, 0.748). The unimodal EHR–GBM achieved AUCs of 0.926, 0.893, and 0.893, in line with the structured-data performance reported in INSPECT (0.848, 0.875, 0.866). Similarly, the bimodal late-fusion model combining EHR–GBM with CT imaging yielded AUCs of 0.863, 0.862, and 0.871, which remain highly comparable to the corresponding INSPECT results (0.867, 0.875, 0.866) despite the smaller cohort used in our study.

These replications indicate that our pipeline faithfully recovers the original behaviour of the unimodal and bimodal models and therefore provides a solid basis for evaluating more advanced multimodal extensions.

Importantly, our Late Fusion models that incorporate radiology reports, a modality absent from the original INSPECT framework, achieve the highest MCC scores across horizons (0.399 for Reports + EHR–GBM at 1-month and 0.479 for Reports + Image + EHR–GBM at 6-month). Both configurations significantly outperform the reproduced baselines as well as the original EHR+CT benchmark ($p < 0.05$). The improvements are especially evident at the 1- and 6-month horizons because descriptive elements in the reports likely capture clinically relevant signals—such as thrombus burden or right-ventricular strain—that are only partially reflected in structured vitals or imaging-derived features.

In summary, although strict numerical comparability with [63] is precluded by differences in data availability and model accessibility, our carefully reproduced baselines and the consistent gains achieved through report-augmented Late Fusion models demonstrate that incorporating narrative radiology information meaningfully extends the prognostic capacity of the INSPECT framework.

Chapter 6

Summary and Conclusions

Lung cancer, and in particular NSCLC, remains one of the most complex and deadly malignancies worldwide, posing major challenges in early diagnosis, risk stratification, and treatment planning. Despite significant advances in imaging, molecular profiling, and targeted therapies, OS rates remain suboptimal. A major limitation of current clinical practice lies in the restricted predictive capacity of conventional prognostic tools, which are typically based on a limited set of structured variables and traditional statistical models that fail to capture the multidimensional and heterogeneous nature of disease progression. There is therefore a critical need for predictive models capable of integrating heterogeneous sources of clinical information to provide accurate, personalized, and clinically meaningful predictions.

Traditional AI-based approaches in oncology often rely predominantly on structured clinical data such as demographics, staging, comorbidities, and laboratory values. While standardized and accessible, these features do not fully represent the richness of clinical reasoning, longitudinal disease evolution, and nuanced patient-specific factors. In contrast, unstructured EHR data, including physician notes and radiology reports, contain essential contextual information regarding symptoms, treatment response, complications, and clinical judgment. In parallel, medical imaging, particularly CT, provides detailed morphological and spatial information about tumor burden and disease extent. However, each modality in isolation offers only a partial representation of the patient.

Building upon these considerations, the central hypothesis of this thesis is that the integration of unstructured EHR data with imaging and structured clinical variables through multimodal deep learning can substantially enhance prognostic modeling in NSCLC and related pulmonary diseases.

This thesis addressed this hypothesis through a sequence of methodological contributions, each corrisponding to a major chapter.

In Chapter 2, we developed a domain-specific NER system tailored to NSCLC, introduc-

ing a novel clinical ontology comprising 25 entity types and fine-tuning a transformer-based architecture (MedBITR3+) on an annotated corpus of over 1,000 sentences. Comparative analyses demonstrated that domain- and language-specific models significantly outperform general-purpose alternatives, achieving an average F1-score of 0.843 compared to 0.817 for multilingual BERT and 0.832 for umBERTo, with statistical significance confirmed by McNemar tests (p near 0). These results highlight the importance of domain adaptation in clinical NER and enable robust semantic structuring of unstructured narratives for downstream prognostic tasks.

Chapter 3 introduced HEAL, a hierarchical attention-based framework designed to derive patient-level embeddings from NER-annotated clinical text for OS and pCR prediction. The hierarchical attention mechanism captures discriminative patterns beyond single-level aggregation strategies, achieving a time-dependent concordance index (C^{td}-index) of 0.639 $\pm$ 0.014 for OS, outperforming clinical-only features (0.590 $\pm$ 0.019) and several ablation models. For pCR prediction, classifiers trained on HEAL-derived representations, particularly XGBoost, achieved a Matthews Correlation Coefficient (MCC) of 0.371 and accuracy of 0.821 when combined with clinical features, demonstrating the complementary value of structured and unstructured information.

In Chapter 4, we proposed a slice-level attention mechanism for CT-based survival prediction using 2D convolutional encoders (EfficientNetB0) with cross-slice aggregation. This approach outperformed 3D CNN baselines both in predictive performance and computational efficiency, achieving a C^{td}-index of 0.584 $\pm$ 0.054 compared to 0.507 $\pm$ 0.074 for ResNet3D. Domain-specific fine-tuning improved external generalization (e.g., from 0.503 to 0.579 on CLARO), and attention maps revealed focus on clinically relevant tumor regions, enhancing interpretability.

Finally, Chapter 5 presented a multimodal fusion framework integrating unstructured text, CT imaging, and structured EHR data for pulmonary embolism mortality prediction using the *INSPECT* dataset. Late fusion strategies emerged as the most robust configuration, achieving MCC improvements of up to +0.141 over unimodal baselines, with statistical significance confirmed through paired t-tests and FDR correction. More complex mechanisms such as cross-attention and armour fusion demonstrated reduced stability, highlighting the importance of balancing architectural sophistication with dataset size and noise sensitivity. These findings emphasize that robust multimodal integration does not necessarily require highly complex fusion schemes.

Collectively, these contributions demonstrate that unstructured EHR data and multimodal integration substantially enhance prognostic modeling. The thesis advances the state of the art by: (i) introducing a lung cancer–specific NER system enabling semantic structur-

ing of clinical narratives; (ii) proposing hierarchical attention for patient-level text representations in survival modeling; (iii) developing slice-level attention for CT-based prognosis; and (iv) evaluating multimodal fusion strategies in pulmonary disease prediction under rigorous statistical validation.

Limitations

Despite these contributions, several limitations must be acknowledged.

First, most experiments rely on retrospective datasets, including single-center cohorts for NSCLC tasks. Although partial external validation was performed for imaging models, comprehensive multi-institutional validation across geographically diverse populations remains limited. Dataset size, particularly for pCR prediction, constrains the complexity of deployable architectures and may limit generalizability.

Second, unstructured EHR data are heterogeneous across institutions in terms of language, formatting, and documentation practices. The manual annotation required for NER development is resource-intensive and may introduce annotation bias. Scalability across healthcare systems therefore requires additional harmonization efforts.

Third, imaging data are affected by variability in acquisition protocols, reconstruction kernels, and scanner types, which can impact robustness. While fine-tuning partially mitigates this issue, domain adaptation strategies require further development.

Fourth, although attention mechanisms provide structural interpretability, attention weights should not be interpreted as causal explanations. More rigorous explainability frameworks are needed to ensure clinical trust and transparency.

Finally, the proposed models were evaluated retrospectively and have not yet undergone prospective validation in real-world clinical workflows. Their impact on decision-making, treatment selection, and patient outcomes has therefore not yet been formally assessed. Regulatory, ethical, and deployment considerations, including model drift, data privacy constraints, and integration within hospital IT systems, remain open challenges.

Future Research Directions

These limitations open several important avenues for future research.

A primary direction involves prospective and multi-center clinical validation studies to evaluate real-world utility, robustness, and clinical impact. Embedding predictive models into clinical decision support systems and assessing their influence on treatment planning and survival outcomes represents a critical next step.

Second, expanding multimodal integration to include genomic, molecular, and radiogenomic features could enhance biological interpretability and enable more precise personalization. Longitudinal modeling of disease trajectories using temporal EHR and imaging sequences may further improve prognostic accuracy.

Third, self-supervised and foundation-model approaches for clinical and multimodal data may reduce reliance on manual annotation and improve representation learning in low-data regimes. Cross-institutional pretraining and federated learning strategies could enhance generalizability while preserving data privacy.

Fourth, advances in uncertainty quantification, calibration, and causal inference are needed to strengthen clinical reliability and trust. Human-in-the-loop systems integrating clinician feedback may facilitate safe deployment.

In conclusion, this thesis demonstrates that leveraging unstructured EHR data and multimodal deep learning significantly enhances prognostic modeling in NSCLC and pulmonary diseases. By combining domain-specific NER, hierarchical attention mechanisms for text, slice-level attention for imaging, and robust multimodal fusion strategies, this work proposes a comprehensive framework for integrating heterogeneous clinical data into predictive modeling. While challenges remain in validation, scalability, and deployment, the presented results provide strong evidence that multimodal AI can meaningfully contribute to the transformation of oncology toward more data-driven, personalized, and effective patient care.

Bibliography

[1] Denise R Aberle, Christine D Berg, WC Black, et al. The national lung screening trial. *overview and study design*, 2011:258, 2007.

[2] Debabrata Acharya and Anirban Mukhopadhyay. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Briefings in functional genomics*, 23(5):549–560, 2024.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[4] Hugo J. W. L. Aerts, Lawrence Wee, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Robin Monshouwer, Benjamin Haibe-Kains, Daniël Rietveld, Frank Hoebers, Marleen M. Rietbergen, C. René Leemans, André Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Data from nsclc-radiomics (version 4) [data set], 2014.

[5] Brad Albertina, Mark Watson, Chandra Holback, and et al. Radiology data from the cancer genome atlas lung adenocarcinoma [tcga-luad] collection. the cancer imaging arch, 2016.

[6] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[7] Krunal D Amin, Elizabeth Hope Weissler, William Ratliff, Alexander E Sullivan, Tara A Holder, Cathleen Bury, Samuel Francis, Brent Jason Theiling, Bradley Hintze, Michael Gao, et al. Development and validation of a natural language processing model to identify low-risk pulmonary embolism in real time to facilitate safe outpatient management. *Annals of Emergency Medicine*, 84(2):118–127, 2024.

[8] Ying An, Xianyun Xia, Xianlai Chen, Fang-Xiang Wu, and Jianxin Wang. Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. *Artificial Intelligence in Medicine*, 127:102282, 2022.

[9] Vatsala Anand, Deepika Koundal, Wael Y Alghamdi, and Bayan M Alsharbi. Smart grading of diabetic retinopathy: an intelligent recommendation-based fine-tuned efficientnetb0 framework. *Frontiers in Artificial Intelligence*, 7:1396160, 2024.

[10] Drahomir Aujesky, D Scott Obrosky, Roslyn A Stone, Thomas E Auble, Arnaud Perrier, Jacques Cornuz, Pierre-Marie Roy, and Michael J Fine. Derivation and validation of a prognostic model for pulmonary embolism. *American journal of respiratory and critical care medicine*, 172(8):1041–1046, 2005.

[11] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Weiruo Zhang, Ann Leung, Michael Kadoch, et al. Data for nsclc radiogenomics collection. the cancer imaging archive, 2017.

[12] Stefano Barco, Luca Valerio, Walter Ageno, Alexander T Cohen, Samuel Z Goldhaber, Beverley J Hunt, Alfonso Iorio, David Jimenez, Frederikus A Klok, Nils Kucher, et al. Age-sex specific pulmonary embolism-related mortality in the usa and canada, 2000–18: an analysis of the who mortality database and of the cdc multiple cause of death database. *The Lancet Respiratory Medicine*, 9(1):33–42, 2021.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[14] Anna Braghetto, Francesca Marturano, Marta Paiusco, Marco Baiesi, and Andrea Bettinelli. Radiomics and deep learning methods for the prediction of 2-year overall survival in lung1 dataset. *Scientific Reports*, 12(1):14132, 2022.

[15] Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431, 2023.

[16] Noa Cahan, Eyal Klang, Edith M Marom, Shelly Soffer, Yiftach Barash, Evyatar Burshtein, Eli Konen, and Hayit Greenspan. Multimodal fusion models for pulmonary embolism mortality prediction. *Scientific Reports*, 13(1):7544, 2023.

[17] Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. A clinical trials corpus annotated with umls entities to enhance

the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19, 2021.

[18] Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors. *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[19] Camillo Maria Caruso, Valerio Guarrasi, Ermanno Cordelli, Rosa Sicilia, Silvia Gentile, Laura Messina, Michele Fiore, Claudia Piccolo, Bruno Beomonte Zobel, Giulio Iannello, et al. A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer. *Journal of Imaging*, 8(11):298, 2022.

[20] Camillo Maria Caruso, Valerio Guarrasi, Sara Ramella, and Paolo Soda. A deep learning approach for overall survival prediction in lung cancer with missing values. *Computer Methods and Programs in Biomedicine*, 254:108308, 2024.

[21] Runsheng Chang, Shouliang Qi, Yanan Wu, Yong Yue, Xiaoye Zhang, and Wei Qian. Nomograms integrating ct radiomic and deep learning signatures to predict overall survival and progression-free survival in nsclc patients treated with chemotherapy. *Cancer Imaging*, 23(1):101, 2023.

[22] Mitchell Chen, Susan J Copley, Patrizia Viola, Haonan Lu, and Eric O Aboagye. Radiomics and artificial intelligence for precision medicine in lung cancer treatment. In *Seminars in cancer biology*, volume 93, pages 97–113. Elsevier, 2023.

[23] Peng Chen, Meng Zhang, Xiaosheng Yu, and Songpu Li. Named entity recognition of chinese electronic medical records based on a hybrid neural network and medical mc-bert. *BMC Medical Informatics and Decision Making*, 22(1):1–13, 2022.

[24] Tao Chen, Jialiang Wen, Xinchen Shen, Jiaqi Shen, Jiajun Deng, Mengmeng Zhao, Long Xu, Chunyan Wu, Bentong Yu, Minglei Yang, et al. Whole slide image based deep learning refines prognosis and therapeutic response evaluation in lung adenocarcinoma. *npj Digital Medicine*, 8(1):69, 2025.

[25] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[26] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[27] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[28] Vedat Cicek, Ahmet Lutfullah Orhan, Faysal Saylik, Vanshali Sharma, Yalcin Tur, Almina Erdem, Mert Babaoglu, Omer Ayten, Solen Taslicukur, Ahmet Oz, et al. Predicting short-term mortality in patients with acute pulmonary embolism with deep learning. *Circulation Journal*, 89(5):602–611, 2025.

[29] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[30] Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

[31] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20, 1995.

[32] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001, 2023.

[33] Mahtab Darvish, Ryan Trask, Patrick Tallon, Mélina Khansari, Lei Ren, Michelle Hershman, and Bardia Yousefi. Ai-enabled lung cancer prognosis. *arXiv preprint arXiv:2402.09476*, 2024.

[34] 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.

[35] Ruining Deng, Nazim Shaikh, Gareth Shannon, and Yao Nie. Cross-modality attention-based multimodal fusion for non-small cell lung cancer (nsclc) patient survival prediction. In *Medical Imaging 2024: Digital and Computational Pathology*, volume 12933, pages 46–50. SPIE, 2024.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[38] Ignacio Diaz-Lorenzo, Alberto Alonso-Burgos, Alfonsa Friera Reyes, Ruben Eduardo Pacios Blanco, Maria del Carmen de Benavides Bernaldo de Quiros, and Guillermo Gallardo Madueño. Current role of ct pulmonary angiography in pulmonary embolism: A state-of-the-art review. *Journal of Imaging*, 10(12):323, 2024.

[39] Pooya Eini, Peyman Eini, Homa Serpoush, Mohammad Rezayee, and Jason Tremblay. Advancing mortality prediction in pulmonary embolism using machine learning algorithms—systematic review and meta-analysis. *Pulmonary Circulation*, 15(3):e70166, 2025.

[40] Nuruzzaman Faruqui, Mohammad Abu Yousuf, Md Whaiduzzaman, AKM Azad, Alistair Barros, and Mohammad Ali Moni. Lungnet: A hybrid deep-cnn model for lung cancer diagnosis using ct and wearable sensor-based medical iot data. *Computers in Biology and Medicine*, 139:104961, 2021.

[41] Alfred P Fishman. *Update: pulmonary diseases and disorders*. McGraw-Hill, 1992.

[42] Patrick M Forde, Jonathan Spicer, Shun Lu, Mariano Provencio, Tetsuya Mitsudomi, Mark M Awad, Enriqueta Felip, Stephen R Broderick, Julie R Brahmer, Scott J Swanson, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. *New England Journal of Medicine*, 386(21):1973–1985, 2022.

[43] Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors. *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[44] Leena K Gautam. Natural language processing-based structured data extraction from unstructured clinical notes. *Notes*, 6:9, 2024.

[45] Olya Grove, Anders E Berglund, Matthew B Schabath, Hugo JWL Aerts, Andre Dekker, Hua Wang, Emmanuel Rios Velazquez, Philippe Lambin, Yuhua Gu, Yoganand Balagurunathan, et al. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PloS one*, 10(3):e0118261, 2015.

[46] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing*, page 105509, 2025.

[47] Zhaoxin Guo, Zhipeng Wang, Ruiquan Ge, Jianxun Yu, Feiwei Qin, Yuan Tian, Yuqing Peng, Yonghong Li, and Changmiao Wang. Pe-mvcnet: Multi-view and cross-modal fusion network for pulmonary embolism prediction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[48] Christoph Haarburger, Philippe Weitz, Oliver Rippel, and Dorit Merhof. Image-based survival prediction for lung cancer patients using cnns. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 1197–1201. IEEE, 2019.

[49] Dan Han, Hao Li, Xin Zheng, Shenbo Fu, Ran Wei, Qian Zhao, Chengxin Liu, Zhongtang Wang, Wei Huang, and Shaoyu Hao. Whole slide image-based weakly supervised deep learning for predicting major pathological response in non-small cell lung cancer following neoadjuvant chemoimmunotherapy: a multicenter, retrospective, cohort study. *Frontiers in Immunology*, 15:1453232, 2024.

[50] Per Christian Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[51] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[52] Syed Moin Hassan, Ruben Mylvaganam, Tekle Didebulidze, Malaika Khalid, Pietro Nardelli, Gregory Piazza, Ruben San Jose Estepar, Michael J Cuttica, Shelsey Johnson, Nam Dao, et al. Leveraging hybrid natural language processing techniques for large-scale pulmonary embolism identification. *JACC: Advances*, page 101845, 2025.

[53] 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.

[54] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[55] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[56] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[57] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4:1–13, 2020.

[58] Ahmed Hosny, Chintan Parmar, Thibaud P Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J Gillies, Raymond H Mak, and Hugo JWL Aerts. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS medicine*, 15(11):e1002711, 2018.

[59] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.

[60] Yuting Hu and Suzan Verberne. Named entity recognition for chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637, 2020.

[61] Peiying Hua, Andrea Olofson, Faraz Farhadi, Liesbeth Hondelink, Gregory Tsongalis, Konstantin Dragnev, Dagmar Hoegemann Savellano, Arief Suriawinata, Laura Tafe, and Saeed Hassanpour. Predicting targeted therapy resistance in non-small cell lung cancer using multimodal machine learning. *arXiv preprint arXiv:2503.24165*, 2025.

[62] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[63] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*, 2023.

[64] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020.

[65] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P Lungren. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147, 2020.

[66] National Cancer Institute. Cancer statistics. https://www.cancer.gov/about-cancer/understanding/statistics, 2022. Accessed: 2025-06-25.

[67] Farhad Islami, Lindsey A Torre, and Ahmedin Jemal. Global trends of lung cancer mortality and smoking prevalence. *Translational lung cancer research*, 4(4):327, 2015.

[68] Bassem Jandoubi and Moulay A Akhloufi. Multimodal artificial intelligence in medical diagnostics. *Information*, 16(7):591, 2025.

[69] David Jiménez, Drahomir Aujesky, Lisa Moores, Vicente Gómez, José Luis Lobo, Fernando Uresandi, Remedios Otero, Manuel Monreal, Alfonso Muriel, Roger D Yusen, et al. Simplification of the pulmonary embolism severity index for prognostication in patients with acute symptomatic pulmonary embolism. *Archives of internal medicine*, 170(15):1383–1389, 2010.

[70] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[71] Aleksandar Kaplar, Milan Stošović, Aleksandra Kaplar, Voin Brković, Radomir Naumović, and Aleksandar Kovačević. Evaluation of clinical named entity recognition methods for serbian electronic health records. *International Journal of Medical Informatics*, 164:104805, 2022.

[72] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[73] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[74] Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka, and Kamalakar Karlapalem. Biomedical ner using novel schema and distant supervision. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 155–160, 2022.

[75] Shanah Kirk, Y. Lee, Prasanna Kumar, Joe Filippini, Brad Albertina, M. Watson, Kimberly Rieger-Christ, and John Lemmerman. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc) (version 4) [data set]. the cancer imaging archive., 2016.

[76] Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025, 2021.

[77] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine*, 5(1):171, 2022.

[78] Veysel Kocaman and David Talby. Accurate clinical and biomedical named entity recognition at scale. *Software Impacts*, 13:100373, 2022.

[79] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

[80] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29, 2017.

[81] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[82] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[83] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[84] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[85] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[86] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[87] Hyun Woo Lee, Young Sik Park, Sangshin Park, and Chang-Hoon Lee. Poor prognosis of nsclc located in lower lobe is partly mediated by lower frequency of egfr mutations. *Scientific reports*, 10(1):14933, 2020.

[88] Butuo Li, Linlin Yang, Huan Zhang, Haoqian Li, Chao Jiang, Yueyuan Yao, Shuping Cheng, Bing Zou, Bingjie Fan, Taotao Dong, et al. Outcome-supervised deep learning on pathologic whole slide images for survival prediction of immunotherapy in patients with non–small cell lung cancer. *Modern Pathology*, 36(8):100208, 2023.

[89] Chengtai Li, Yiming Zhang, Ying Weng, Boding Wang, and Zhenzhu Li. Natural language processing applications for computer-aided diagnosis in oncology. *Diagnostics*, 13(2):286, 2023.

[90] Juanzi Li, Ming Zhou, Guilin Qi, Ni Lao, Tong Ruan, and Jianfeng Du. *Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence: Second China Conference, CCKS 2017, Chengdu, China, August 26–29, 2017, Revised Selected Papers*, volume 784. Springer, 2018.

[91] Xia Li, Qinghua Wen, Hu Lin, Zengtao Jiao, and Jiangtao Zhang. Overview of ccks 2020 task 3: named entity recognition and event extraction in chinese electronic medical records. *Data Intelligence*, 3(3):376–388, 2021.

[92] Yan-Jun Li, Hsin-Hung Chou, Peng-Chan Lin, Meng-Ru Shen, and Sun-Yuan Hsieh. A novel deep learning-based algorithm combining histopathological features with tissue areas to predict colorectal cancer survival from whole-slide images. *Journal of Translational Medicine*, 21(1):731, 2023.

[93] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[94] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[95] William Adam Libling, Ronald Korn, and Glen J Weiss. Review of the use of radiomics to assess the risk of recurrence in early-stage non-small cell lung cancer. *Translational Lung Cancer Research*, 12(7):1575, 2023.

[96] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[97] Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of biomedical informatics*, 145:104466, 2023.

[98] Junwei Liu, Xiaoping Cen, Chenxin Yi, Feng-ao Wang, Junxiang Ding, Jinyu Cheng, Qinhua Wu, Baowen Gai, Yiwen Zhou, Ruikun He, et al. Challenges in ai-driven biomedical multimodal data fusion and analysis. *Genomics, Proteomics & Bioinformatics*, 23(1):qzaf011, 2025.

[99] Filippo Lococo, Galal Ghaly, Marco Chiappetta, Sara Flamini, Jessica Evangelista, Emilio Bria, Alessio Stefani, Emanuele Vita, Antonella Martino, Luca Boldrini, et al. Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: a narrative review. *Cancers*, 16(10):1832, 2024.

[100] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[101] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, page 100216, 2023.

[102] Aditya Chandra Mandal and Abhijeet Phatak. Optimizing deep learning based retinal diseases classification on optical coherence tomography scans. In *European Conference on Biomedical Optics*, page 1263220. Optica Publishing Group, 2023.

[103] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791, 2023.

[104] Hoda Memarzadeh, Nasser Ghadiri, and Maryam Lotfi Shahreza. Assessing mortality prediction through different representation models based on concepts extracted from clinical notes. *arXiv preprint arXiv:2207.10872*, 2022.

[105] S Mithun, Ashish Kumar Jha, Umesh B Sherkhane, Vinay Jaiswar, Nilendu C Purandare, V Rangarajan, A Dekker, Sander Puts, Inigo Bermejo, and L Wee. Development and validation of deep learning and bert models for classification of lung cancer radiology reports. *Informatics in Medicine Unlocked*, page 101294, 2023.

[106] Proportional Hazards Model. Cox proportional hazards model. 2021.

[107] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981, 2022.

[108] Gustav Müller-Franzes, Firas Khader, Robert Siepmann, Tianyu Han, Jakob Nikolas Kather, Sven Nebelung, and Daniel Truhn. Medical slice transformer for improved diagnosis and explainability on 3d medical images with dinov2. *Scientific Reports*, 15(1):23979, 2025.

[109] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[110] Sankaran Narayanan, Kaivalya Mannam, Pradeep Achan, Maneesha V Ramesh, P Venkat Rangan, and Sreeranga P Rajan. A contextual multi-task neural approach to medication and adverse events identification from clinical text. *Journal of biomedical informatics*, 125:103960, 2022.

[111] David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, page 105122, 2023.

[112] Tatsuya Nishikawa, Takeshi Fujita, Toshitaka Morishima, Sumiyo Okawa, Terutaka Hino, Taku Yasui, Wataru Shioyama, Toru Oka, Isao Miyashiro, and Masashi Fujita. Prognostic effect of incidental pulmonary embolism on long-term mortality in cancer patients. *Circulation Journal*, 88(2):198–204, 2024.

[113] Kevin O'Gallagher, Jack Wu, and Richard JB Dobson. Calming the storm: Natural language processing applied to pulmonary embolism research, 2025.

[114] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[115] Hafiza Padinharayil, Jinsu Varghese, Mithun Chacko John, Golgodu Krishnamurthy Rajanikant, Cornelia M Wilson, Minnatallah Al-Yozbaki, Kaviyarasi Renu, Saikat Dewanjee, Rupa Sanyal, Abhijit Dey, et al. Non-small cell lung carcinoma (nsclc): Implications on molecular pathology and advances in early diagnostics and therapeutics. *Genes & Diseases*, 10(3):960–989, 2023.

[116] Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H Mak, Nicolai J Birkbak, et al. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, 6(3):354–367, 2024.

[117] Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

[118] Domenico Paolo, Carlo Greco, Alessio Cortellini, Sara Ramella, Paolo Soda, Alessandro Bria, and Rosa Sicilia. Hierarchical embedding attention for overall survival prediction in lung cancer from unstructured ehrs. *BMC Medical Informatics and Decision Making*, 25(1):169, 2025.

[119] Loreto Parisi, Simone Francia, and Paolo Magnani. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto, 2020.

[120] Min-Koo Park, Jin-Muk Lim, Jinwoo Jeong, Yeongjae Jang, Ji-Won Lee, Jeong-Chan Lee, Hyungyu Kim, Euiyul Koh, Sung-Joo Hwang, Hong-Gee Kim, et al. Deep-learning

algorithm and concomitant biomarker identification for nsclc prediction using multi-omics data integration. *Biomolecules*, 12(12):1839, 2022.

[121] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[122] Charmy H Patel, Devang Undaviya, Harsh Dave, Sheshang Degadwala, and Dhairya Vyas. Efficientnetb0 for brain stroke classification on computed tomography scan. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 713–718. IEEE, 2023.

[123] Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527, 2010.

[124] Uyen Phan and Nhung Nguyen. Simple semantic-based data augmentation for named entity recognition in biomedical texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 123–129, 2022.

[125] Uyen Phan, Phuong NV Nguyen, and Nhung Nguyen. A named entity recognition corpus for vietnamese biomedical texts to support tuberculosis treatment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3601–3609, 2022.

[126] 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.

[127] Sara Ramella, Michele Fiore, Carlo Greco, Ermanno Cordelli, Rosa Sicilia, Mario Merone, Elisabetta Molfese, Marianna Miele, Patrizia Cornacchione, Edy Ippolito, et al. A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients. *PloS one*, 13(11):e0207455, 2018.

[128] Sina Rashedi, Darsiya Krishnathasan, Candrika Khairani, Antoine Bejjani, Ying-Chih Lo, Mehrdad Zarghami, Shiwani Mahajan, Cesar Caraballo, jose victor jimenez ceja, David Jimenez, et al. Accuracy of rule-based natural language processing models for identification of pulmonary embolism. *Circulation*, 150(Suppl_1):A4147439–A4147439, 2024.

[129] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[130] Steven J Rigatti. Random forest. *Journal of insurance medicine*, 47(1):31–39, 2017.

[131] Seyed-Ali Sadegh-Zadeh, Hanie Sakha, Sobhan Movahedi, Aniseh Fasihi Harandi, Samad Ghaffari, Elnaz Javanshir, Syed Ahsan Ali, Zahra Hooshanginezhad, and Reza Hajizadeh. Advancing prognostic precision in pulmonary embolism: a clinical and laboratory-based artificial intelligence approach for enhanced early mortality risk stratification. *Computers in Biology and Medicine*, 167:107696, 2023.

[132] 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.

[133] Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language resources and evaluation*, 50:523–548, 2016.

[134] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

[135] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17–48, 2023.

[136] Benjamin D Simon, Kutsev Bengisu Ozyoruk, David G Gelikman, Stephanie A Harmon, and Barış Türkbey. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, 31(4):303, 2025.

[137] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[138] OSWALDO SOLARTE PABÓN, Orlando Montenegro, Alvaro García, Alejandro Rodriguez-Gonzalez, Juan Cristobal Sanchez, Víctor Robles, Mariano Provencio, and Ernestina Menasalvas. A deep learning approach to extract lung cancer information from spanish clinical texts. *Available at SSRN 4049602*.

[139] Oswaldo Solarte-Pabón, Orlando Montenegro, Alvaro García-Barragán, Maria Torrente, Mariano Provencio, Ernestina Menasalvas, and Víctor Robles. Transformers for extracting breast cancer information from spanish clinical narratives. *Artificial Intelligence in Medicine*, 143:102625, 2023.

[140] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

[141] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

[142] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19, 2015.

[143] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.

[144] Zhaoyi Sun, Mingquan Lin, Qingqing Zhu, Qianqian Xie, Fei Wang, Zhiyong Lu, and Yifan Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of biomedical informatics*, 146:104482, 2023.

[145] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[146] Shan Suthaharan. Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235. Springer, 2016.

[147] Yasasvy Tadepalli, Meenakshi Kollati, Swaraja Kuraparthi, and Padmavathi Kora. Efficientnet-b0 based monocular dense-depth map estimation. *Traitement du Signal*, 38(5), 2021.

[148] Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond struc-

tured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021.

[149] Priyabrata Thatoi, Rohit Choudhary, Ashish Shiwlani, Hamza Ahmed Qureshi, and Sooraj Kumar. Natural language processing (nlp) in the extraction of clinical information from electronic health records (ehrs) for cancer prognosis. *International Journal*, 10(4):2676–2694, 2023.

[150] Felipe Soares Torres, Shazia Akbar, Srinivas Raman, Kazuhiro Yasufuku, Carola Schmidt, Ahmed Hosny, Felix Baldauf-Lenschen, and Natasha B Leighl. End-to-end non–small-cell lung cancer prognostication using deep learning applied to pretreatment computed tomography. *JCO Clinical Cancer Informatics*, 5:1141–1150, 2021.

[151] Matteo Tortora, Ermanno Cordelli, Rosa Sicilia, Marianna Miele, Paolo Matteucci, Giulio Iannello, Sara Ramella, and Paolo Soda. Deep reinforcement learning for fractionated radiotherapy in non-small cell lung carcinoma. *Artificial Intelligence in Medicine*, 119:102137, 2021.

[152] Matteo Tortora, Ermanno Cordelli, Rosa Sicilia, Lorenzo Nibid, Edy Ippolito, Giuseppe Perrone, Sara Ramella, and Paolo Soda. Radiopathomics: Multimodal learning in non-small cell lung cancer for adaptive radiotherapy. *IEEE Access*, 11:47563–47578, 2023.

[153] 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.

[154] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[155] Diem Vuong, Marta Bogowicz, Sarah Denzler, Carol Oliveira, Robert Foerster, Florian Amstutz, Hubert S Gabryś, Jan Unkelbach, Sven Hillinger, Sandra Thierstein, et al. Comparison of robust to standardized ct radiomics models to predict overall survival for non-small cell lung cancer patients. *Medical physics*, 47(9):4045–4053, 2020.

[156] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[157] Yongzhong Wu, Smitha Antony, Jennifer L Meitzler, and James H Doroshow. Molecular mechanisms underlying chronic inflammation-associated cancers. *Cancer letters*, 345(2):164–173, 2014.

[158] Yujiao Wu, Yaxiong Wang, Xiaoshui Huang, Haofei Wang, Fan Yang, Wenwen Sun, Steven W Su, and Sai Ho Ling. Multimodal learning for non-small cell lung cancer prognosis. *Biomedical Signal Processing and Control*, 106:107663, 2025.

[159] Fei Xu, Wen-qiang Cui, Cun Liu, Fubin Feng, Ruijuan Liu, Jingtao Zhang, and Changgang Sun. Prognostic biomarkers correlated with immune infiltration in non-small cell lung cancer. *FEBS Open Bio*, 13(1):72–88, 2023.

[160] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.

[161] Huan Yang, Minglei Yang, Jiani Chen, Guocong Yao, Quan Zou, and Linpei Jia. Multimodal deep learning approaches for precision oncology: a comprehensive review. *Briefings in Bioinformatics*, 26(1):bbae699, 2025.

[162] Xiaorong Yang, Tongchao Zhang, Xiangwei Zhang, Chong Chu, and Shaowei Sang. Global burden of lung cancer attributable to ambient fine particulate matter pollution in 204 countries and territories, 1990–2019. *Environmental research*, 204:112023, 2022.

[163] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020.

[164] Qianyu Yuan, Tianrun Cai, Chuan Hong, Mulong Du, Bruce E Johnson, Michael Lanuti, Tianxi Cai, and David C Christiani. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. *JAMA Network Open*, 4(7):e2114723–e2114723, 2021.

[165] Huanyao Zhang, Danqing Hu, Huilong Duan, Shaolei Li, Nan Wu, and Xudong Lu. A novel deep learning approach to extract chinese clinical entities for lung cancer screening and staging. *BMC Medical Informatics and Decision Making*, 21(2):1–12, 2021.

[166] Jinsong Zhang, Xiaomei Yu, Zhichao Wang, and Xiangwei Zheng. Gwbner: A named entity recognition method based on character glyph and word boundary features for

chinese ehrs. *Journal of King Saud University-Computer and Information Sciences*, 35(8):101654, 2023.

[167] Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*, 2020.

[168] Wei Zhang, Yu Gu, Hao Ma, Lidong Yang, Baohua Zhang, Jing Wang, Meng Chen, Xiaoqi Lu, Jianjun Li, Xin Liu, Dahua Yu, Ying Zhao, Siyuan Tang, and Qun He. A novel multimodal computer-aided diagnostic model for pulmonary embolism based on hybrid transformer-cnn and tabular transformer. *Physical and Engineering Sciences in Medicine*, 48:1107 – 1126, 2025.

[169] Ying Zhang, Baohang Zhou, Kehui Song, Xuhui Sui, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. Pm2f2n: Patient multi-view multi-modal feature fusion networks for clinical outcome prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1985–1994, 2022.

[170] Jun Zhao, Frank Van Harmelen, Jie Tang, Xianpei Han, Quan Wang, and Xianyong Li. *Knowledge Graph and Semantic Computing. Knowledge Computing and Language Understanding: Third China Conference, CCKS 2018, Tianjin, China, August 14–17, 2018, Revised Selected Papers*, volume 957. Springer, 2018.

[171] Sunyi Zheng, Jiapan Guo, Johannes A Langendijk, Stefan Both, Raymond NJ Veldhuis, Matthijs Oudkerk, Peter MA van Ooijen, Robin Wijsman, and Nanna M Sijtsema. Survival prediction for stage i-iiia non-small cell lung cancer using deep learning. *Radiotherapy and oncology*, 180:109483, 2023.

[172] Zhusi Zhong, Helen Zhang, Fayez H Fayad, Andrew C Lancaster, John Sollee, Shreyas Kulkarni, Cheng Ting Lin, Jie Li, Xinbo Gao, Scott Collins, et al. Pulmonary embolism mortality prediction using multimodal learning based on computed tomography angiography and clinical data. *arXiv preprint arXiv:2406.01302*, 2024.

[173] Lina Zhou, Chenkai Mao, Tingting Fu, Xiao Ding, Luca Bertolaccini, Ao Liu, Junjun Zhang, and Shicheng Li. Development of an ai model for predicting hypoxia status and prognosis in non-small cell lung cancer using multi-modal data. *Translational Lung Cancer Research*, 13(12):3642, 2024.

[174] Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from

electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 2022.

[175] Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

[176] Xiaoyan Zhu, Bing Qin, Xiaodan Zhu, Ming Liu, and Longhua Qian. *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019, Revised Selected Papers*, volume 1134. Springer Nature, 2020.

[177] Marco Zuin, Anju Nohria, Stanislav Henkin, Darsiya Krishnathasan, Alyssa Sato, and Gregory Piazza. Pulmonary embolism–related mortality in patients with cancer. *JAMA Network Open*, 8(2):e2460315–e2460315, 2025.

Appendices

Appendix A

Supplementary Results for Multimodal Pulmonary Embolism Prediction

A.1 Complete Unimodal Results

Table A.1: Performance of unimodal models across 1-, 6-, and 12-month mortality prediction tasks. Results are reported as mean ± standard deviation over five runs. Best MCC values for each time horizon are highlighted in **bold**

Model	1-Month			6-Month			12-Month		
	MCC	AUC	AUPRC	MCC	AUC	AUPRC	MCC	AUC	AUPRC
EHR-GBM	0.258 ± 0.059	0.926 ± 0.004	0.429 ± 0.019	**0.367 ± 0.072**	0.893 ± 0.008	0.503 ± 0.021	**0.454 ± 0.023**	0.893 ± 0.002	0.561 ± 0.007
EHR-AE	**0.269 ± 0.013**	0.790 ± 0.028	0.200 ± 0.014	0.352 ± 0.016	0.817 ± 0.014	0.354 ± 0.013	0.395 ± 0.008	0.812 ± 0.013	0.433 ± 0.012
Report	0.242 ± 0.008	0.837 ± 0.005	0.210 ± 0.007	0.329 ± 0.030	0.825 ± 0.003	0.402 ± 0.008	0.354 ± 0.014	0.821 ± 0.007	0.481 ± 0.012
Image	0.221 ± 0.010	0.800 ± 0.003	0.173 ± 0.007	0.297 ± 0.016	0.790 ± 0.008	0.323 ± 0.010	0.304 ± 0.028	0.781 ± 0.013	0.363 ± 0.013

A.2 Complete Multimodal Results

Table A.2: Performance comparison of unimodal and multimodal models across 1-, 6-, and 12-month mortality prediction tasks. The table reports MCC, AUC, and AUPRC for each modality and fusion scheme. Results are reported as mean ± standard deviation over five runs. Best MCC values for each time horizon are highlighted in **bold**.

Model	1-Month			6-Month			12-Month		
	MCC	AUC	AUPRC	MCC	AUC	AUPRC	MCC	AUC	AUPRC
Unimodal									
EHR-GBM	0.258 ± 0.059	0.926 ± 0.004	0.429 ± 0.019	0.367 ± 0.072	0.893 ± 0.008	0.503 ± 0.021	0.454 ± 0.023	0.893 ± 0.002	0.561 ± 0.007
EHR-AE	0.269 ± 0.013	0.790 ± 0.028	0.200 ± 0.014	0.352 ± 0.016	0.817 ± 0.014	0.354 ± 0.013	0.395 ± 0.008	0.812 ± 0.013	0.433 ± 0.012
Report	0.242 ± 0.008	0.837 ± 0.005	0.210 ± 0.007	0.329 ± 0.030	0.825 ± 0.003	0.402 ± 0.008	0.354 ± 0.014	0.821 ± 0.007	0.481 ± 0.012
Image	0.221 ± 0.010	0.800 ± 0.003	0.173 ± 0.007	0.297 ± 0.016	0.790 ± 0.008	0.323 ± 0.010	0.304 ± 0.028	0.781 ± 0.013	0.363 ± 0.013
Early Fusion									
EHR-AE, Report	0.296 ± 0.014	0.845 ± 0.005	0.228 ± 0.011	0.371 ± 0.018	0.808 ± 0.002	0.365 ± 0.003	0.407 ± 0.009	0.795 ± 0.003	0.424 ± 0.007
EHR-AE, Image	0.302 ± 0.011	0.842 ± 0.005	0.234 ± 0.005	0.380 ± 0.016	0.810 ± 0.004	0.369 ± 0.006	0.405 ± 0.010	0.794 ± 0.003	0.423 ± 0.003
EHR-TSVD, Report	0.280 ± 0.027	0.846 ± 0.005	0.245 ± 0.013	0.392 ± 0.006	0.856 ± 0.009	0.434 ± 0.014	0.419 ± 0.010	0.858 ± 0.005	0.490 ± 0.017
EHR-TSVD, Image	0.274 ± 0.013	0.842 ± 0.012	0.212 ± 0.024	0.376 ± 0.011	0.849 ± 0.011	0.402 ± 0.018	0.390 ± 0.017	0.847 ± 0.004	0.477 ± 0.018
Report, Image	0.244 ± 0.024	0.841 ± 0.001	0.231 ± 0.009	0.350 ± 0.016	0.829 ± 0.001	0.397 ± 0.003	0.366 ± 0.012	0.828 ± 0.004	0.472 ± 0.007
EHR-AE, Report, Image	0.302 ± 0.008	0.845 ± 0.006	0.237 ± 0.014	0.378 ± 0.014	0.810 ± 0.002	0.369 ± 0.003	0.398 ± 0.007	0.790 ± 0.006	0.426 ± 0.008
EHR-TSVD, Report, Image	0.293 ± 0.009	0.861 ± 0.006	0.239 ± 0.006	0.393 ± 0.008	0.857 ± 0.005	0.419 ± 0.007	0.426 ± 0.008	0.853 ± 0.008	0.478 ± 0.021
Late Fusion									
Reports, EHR-AE	0.286 ± 0.015	0.865 ± 0.005	0.250 ± 0.011	0.388 ± 0.019	0.861 ± 0.005	0.427 ± 0.011	0.424 ± 0.005	0.860 ± 0.005	0.519 ± 0.014
Reports, EHR-GBM	**0.399 ± 0.050**	0.884 ± 0.008	0.394 ± 0.013	0.472 ± 0.011	0.872 ± 0.008	0.473 ± 0.011	0.494 ± 0.008	0.875 ± 0.004	0.551 ± 0.007
Image, EHR-AE	0.297 ± 0.017	0.846 ± 0.011	0.268 ± 0.013	0.395 ± 0.016	0.855 ± 0.005	0.442 ± 0.009	0.426 ± 0.004	0.852 ± 0.006	0.506 ± 0.009
Image, EHR-GBM	0.374 ± 0.025	0.863 ± 0.012	0.372 ± 0.019	0.467 ± 0.031	0.862 ± 0.011	0.500 ± 0.022	**0.497 ± 0.011**	0.871 ± 0.002	0.574 ± 0.011
Reports, Image	0.275 ± 0.015	0.855 ± 0.004	0.232 ± 0.005	0.375 ± 0.011	0.848 ± 0.003	0.431 ± 0.008	0.398 ± 0.012	0.841 ± 0.003	0.502 ± 0.007
Reports, Image, EHR-AE	0.331 ± 0.014	0.875 ± 0.005	0.285 ± 0.012	0.409 ± 0.014	0.874 ± 0.002	0.477 ± 0.010	0.444 ± 0.008	0.871 ± 0.004	0.553 ± 0.010
Reports, Image, EHR-GBM	0.362 ± 0.016	0.884 ± 0.007	0.384 ± 0.016	**0.479 ± 0.011**	0.878 ± 0.006	0.508 ± 0.015	0.488 ± 0.002	0.879 ± 0.003	0.581 ± 0.009
Cross Fusion									
EHR-TSVD → Image → Report	0.240 ± 0.023	0.813 ± 0.004	0.189 ± 0.009	0.321 ± 0.017	0.811 ± 0.004	0.365 ± 0.012	0.351 ± 0.028	0.807 ± 0.004	0.447 ± 0.007
EHR-AE → Image → Report	0.227 ± 0.011	0.808 ± 0.010	0.190 ± 0.019	0.327 ± 0.019	0.807 ± 0.014	0.351 ± 0.014	0.404 ± 0.022	0.826 ± 0.015	0.468 ± 0.010
EHR-TSVD → Report	0.247 ± 0.021	0.842 ± 0.005	0.245 ± 0.027	0.350 ± 0.028	0.828 ± 0.010	0.389 ± 0.009	0.377 ± 0.035	0.831 ± 0.007	0.470 ± 0.015
EHR-TSVD → Image	0.181 ± 0.024	0.772 ± 0.031	0.151 ± 0.021	0.294 ± 0.063	0.798 ± 0.021	0.308 ± 0.048	0.294 ± 0.119	0.801 ± 0.015	0.338 ± 0.051
EHR-AE → Report	0.285 ± 0.032	0.845 ± 0.011	0.234 ± 0.018	0.363 ± 0.019	0.811 ± 0.005	0.354 ± 0.015	0.391 ± 0.039	0.801 ± 0.015	0.434 ± 0.024
Image → Report	0.236 ± 0.025	0.823 ± 0.008	0.189 ± 0.018	0.341 ± 0.016	0.820 ± 0.006	0.369 ± 0.014	0.361 ± 0.013	0.818 ± 0.002	0.459 ± 0.006
EHR-TSVD → Report → Image	0.084 ± 0.032	0.651 ± 0.016	0.093 ± 0.007	0.146 ± 0.052	0.623 ± 0.012	0.166 ± 0.011	0.087 ± 0.051	0.630 ± 0.018	0.210 ± 0.018
EHR-AE → Report → Image	0.124 ± 0.083	0.685 ± 0.065	0.111 ± 0.031	0.204 ± 0.158	0.704 ± 0.092	0.229 ± 0.071	0.165 ± 0.184	0.669 ± 0.089	0.251 ± 0.074
Image → Report → EHR-AE	0.305 ± 0.007	0.835 ± 0.005	0.214 ± 0.011	0.372 ± 0.023	0.810 ± 0.007	0.352 ± 0.024	0.397 ± 0.010	0.794 ± 0.008	0.415 ± 0.010
Image → Report → EHR-TSVD	0.250 ± 0.021	0.813 ± 0.017	0.167 ± 0.031	0.346 ± 0.017	0.822 ± 0.005	0.333 ± 0.032	0.388 ± 0.013	0.831 ± 0.004	0.426 ± 0.009
Armour Fusion									
EHR-TSVD, Image, Report	0.239 ± 0.027	0.801 ± 0.018	0.182 ± 0.020	0.360 ± 0.030	0.815 ± 0.013	0.341 ± 0.022	0.391 ± 0.024	0.820 ± 0.009	0.406 ± 0.023
EHR-AE, Image, Report	0.284 ± 0.023	0.821 ± 0.006	0.191 ± 0.007	0.372 ± 0.015	0.783 ± 0.008	0.341 ± 0.010	0.420 ± 0.012	0.782 ± 0.007	0.415 ± 0.002

Appendix B

Implementation Details

This appendix provides a consolidated overview of the main implementation details, preprocessing strategies, model configurations, and training settings adopted across the different experimental chapters of this thesis. The purpose of this section is to improve transparency and reproducibility by collecting technical details that are otherwise distributed across chapters.

B.1 Named Entity Recognition (Chapter 2)

B.1.1 Text Preprocessing

- Removal of non-informative formatting artifacts
- Sentence segmentation using clinical-aware tokenizer
- Token-level alignment for entity span consistency
- BIO tagging scheme

B.1.2 Transformer Backbone

- Model: biomedical pretrained BERT
- Maximum sequence length: 512
- Tokenizer: WordPiece

B.1.3 Training Configuration

- Optimizer: Adam
- Learning rate: 5×10^5
- Batch size: 8
- Number of epochs: 12
- Gradient clipping: No
- Hardware: NVIDIA GPU T4

B.1.4 Evaluation Strategy

- 10-fold cross-validation
- Entity-level micro and macro F1-score
- Strict span matching

B.2 Survival and Treatment Response Models (Chapter 3)

B.2.1 Text Representation (HEAL Architecture)

- Token embedding model: NER pretrained biomedical BERT (only NER tokens)
- Hierarchical attention:
  - Token-level attention dimension: [num sentences, num tokens, 768]
  - Sentence-level attention dimension: [num sentences, 768]
- Output embedding size: 768

B.2.2 Survival Modeling

- Survival framework: DeepHit-based modeling
- Loss function: Log-likelihood + ranking loss

- Time discretization bins
- Evaluation metric: Time-dependent concordance index (C^{td}-index)

B.2.3 pCR Prediction

- Classifier: Multi-layer perceptron or Machine Learning model
- Loss function: Binary Cross-Entropy
- Evaluation metric: AUC

B.2.4 Training Strategy

- Data split: patient-level separation
- Cross-validation: 10-fold

B.3 Imaging-Based Survival Prediction (Chapter 4)

B.3.1 CT Preprocessing

- Intensity clipping: [-1000, 400] HU
- Normalization: z-score
- Slice selection strategy: sampled slices (only slices with lungs)
- Lungs cropping along z axis
- Resize resolution: 224×224

B.3.2 CNN Backbone

- Architecture: EfficientNetB0
- Pretrained weights: ImageNet
- Fine-tuning strategy: fine-tuning of the last layer

B.3.3 Training Configuration

- Optimizer: AdamW
- Learning rate: 1×10^{-4}
- Batch size: 4
- Epochs: 100
- Data augmentation: (flip and rotation within a range of 20°)
- Evaluation: 10-fold CV
- Metric: C^{td}-index

B.4 Multimodal Fusion Models (Chapter 5)

Dataset: INSPECT dataset

B.4.1 Modalities

- Structured EHR features
- CTPA imaging
- Radiology reports

B.4.2 Text Model

- Transformer backbone: Clinical-Longformer
- Max token length: 4096 tokens
- Fine-tuned: No

B.4.3 Imaging Model

- Architecture: ResNetV2
- Pretraining: ImageNet
- Feature embedding size: 6145

B.4.4 Fusion Strategies

- Early fusion: concatenation layer size 7041/7681
- Late fusion: decision-level averaging
- Cross-attention:
  - Attention heads: 8
  - Hidden dimension: 512

B.4.5 Training Settings

- Split strategy: official patient-level split
- Loss: Cross-Entropy
- Imbalance handling: class weights
- Metrics: MCC, AUC, AUPRC

B.5 Software and Environment

- Python version: 3.11.3
- Framework: PyTorch 2.1.2
- CUDA: 12.1.1
- Transformers library version: 4.40.2
- Random seed control: Enabled
- Deterministic training: Yes