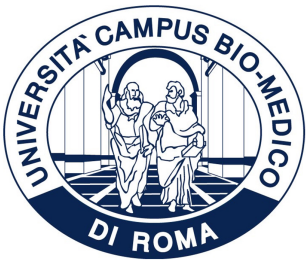


Università Campus Bio-Medico di Roma

Faculty of Engineering
PhD Program in Science and Engineering for Humans and the Environment
XXXVII Ciclo
Curriculum in Information technology in biomedicine

Human-Centred Design of Low-Impact Sensing and Affective Technologies in Healthcare

Author: Costanza Cenerini
Supervisors: Prof. Giorgio Pennazza
Prof. Flavio Keller
Co-supervisor: Prof. Luca Vollero

April 04, 2025

Al mio babbo,
che mi ha insegnato ad ascoltare la musica e a percepire la bellezza
nell'arte, seminando inconsapevolmente i semi di questa ricerca.

Contents

Chapter 1

Introduction

M.C. Escher, "Relativity" (1953), lithograph, 29.4 × 28.2 cm

1.1 Background and Motivation

In an era where technology increasingly permeates every aspect of human life, the way we interact with these technologies has become a crucial determinant of their success and adoption. While technological advancement often focuses on improving functionality and performance, the human experience of using these technologies - their intrusiveness, their demands on our attention, and their integration into our daily routines - has emerged as an equally critical consideration [1, 2]. This is particularly evident in healthcare and assistive technologies, where the effectiveness of a solution must be balanced against its impact on the user's quality of life [3, 4].

Traditional approaches to health monitoring and assistive technologies often prioritize accuracy and reliability over user experience. Consider continuous glucose monitoring systems that require frequent calibration through finger pricks [5], or sensory substitution devices that demand significant cognitive effort from users [6]. While these solutions provide valuable functionality, their intrusive nature can lead to reduced compliance, user fatigue, and ultimately, diminished effectiveness. The challenge lies not just in creating functional technologies, but in developing solutions that work harmoniously with natural human behaviour and perception [7, 8].

Recent advances in sensing technology, artificial intelligence, and our understanding of human perception have opened new possibilities for more non-invasive approaches. Emerging research in affective computing demonstrates that technologies can be made more intuitive by incorporating emotional awareness [9, 10]. Studies in crossmodal perception reveal how different sensory modalities naturally interact [11], suggesting paths toward more organic sensory substitution systems. However, these advances often develop in isolation, missing opportunities for synergistic integration.

The need for low-impact solutions extends beyond mere user comfort. When technologies demand less conscious attention and adapt more naturally to human behaviour, they can achieve better outcomes. This has been demonstrated in various contexts, from healthcare monitoring to assistive

devices for sensory impairment [12, 13].

Understanding human perception and emotion plays a crucial role in developing truly non-invasive solutions. The way humans naturally process and respond to sensory information, including the critical role of emotional processing, must inform the design of these technologies [9, 10]. The effectiveness of unobtrusive approaches has been demonstrated across multiple domains, this understanding enables the development of systems that not only minimize physical intrusion but also reduce cognitive load and emotional burden on users.

Despite these advances and insights, significant gaps remain in our ability to create truly low-impact technological solutions. Integration of emotional awareness in sensory substitution [14] and creation of natural human-machine interfaces all present ongoing challenges. This thesis addresses these gaps through a multifaceted approach that combines theoretical understanding of human perception with practical applications in sensing and interface design.

This research aims to advance the field of unobtrusive technologies through several interconnected investigations. By examining the role of emotion in crossmodal perception, developing new approaches to affect-aware computing, and creating novel sensing solutions, this work contributes to the broader goal of developing technologies that enhance human capabilities while minimizing their impact on natural behaviour and experience.

1.2 Research Objectives

This thesis aims to advance the development of unobtrusive technological solutions that enhance quality of life while maintaining natural human interaction patterns. The research is driven by the hypothesis that understanding and incorporating human emotional and perceptual processes can lead to more effective and acceptable solutions in both healthcare monitoring and sensory assistance.

Figure 1.1. Research framework outlining the three main research areas investigated in this thesis: understanding human perception and emotion (Chapter 2), developing emotion-aware technologies (Chapter 3), and advancing non-invasive sensing (Chapter 4). Each area contributes to the central hypothesis that human-centred design approaches to sensing and affective technologies can improve user acceptance and efficacy in healthcare applications.

As shown in Figure 1.1, each research area directly addresses specific gaps identified in the current literature. The investigation of crossmodal perception and emotional mediation (Chapter 2, green) extends beyond existing work by examining age-related differences and quantifying emotional influence in sensory integration. The development of emotion-aware technologies (Chapter 3, light blue) advances the state of the art through standardized validation methodologies for FER algorithms and novel applications in sensory substitution and therapeutic contexts. Finally, the advancement of non-invasive sensing (Chapter 4, pink) pushes beyond current approaches by fundamentally reconsidering calibration strategies, measurement protocols, and analysis speed to enhance user experience while maintaining technical performance. Together, these research areas create a comprehensive framework that extends significantly beyond the current state of the art in human-centred healthcare technologies.

Primary Objectives

The primary objectives of this research are:

1. **Understanding Human Perception and Emotion**
    - Investigate the role of emotion in crossmodal perception
    - Develop frameworks for quantifying emotional responses
    - Establish the relationship between emotion and user acceptance
2. **Developing Emotion-Aware Technologies**
    - Create systems that adapt to and work with human emotional processes
    - Establish methodologies for validating emotion recognition systems
    - Design interfaces that reduce cognitive and emotional burden
3. **Advancing Non-Invasive Sensing Solutions**

- Develop sensing approaches that minimize user intervention
- optimise calibration strategies to reduce system invasiveness
- Explore alternative measurement techniques that prioritize user comfort

Scope and Impact

The scope of this research encompasses theoretical investigation, system development, and practical validation. While the specific applications range from sensory substitution to healthcare monitoring, they converge on the common goal of developing more non-invasive solutions. The research explicitly considers both the technical performance of these solutions and their impact on user experience, recognising that success requires excellence in both dimensions.

This work contributes to multiple fields, including human-computer interaction, assistive technology, and healthcare monitoring. By focusing on the development of unobtrusive solutions, it addresses a critical need in these domains while advancing our understanding of how to create technologies that work in harmony with natural human processes.

1.3 Thesis Structure

This thesis is organised into five chapters, followed by two appendices that provide complementary research in related domains:

Main Chapters

Chapter 2: Crossmodal Perception and Emotion
Establishes the theoretical foundation with the aim of understanding how humans integrate information across different sensory modalities. Presents original research on the role of emotion in audiovisual associations, with particular focus on its implications for developing more natural human-machine interfaces.

Chapter 3: Affective Computing and Emotion Recognition
Explores the translation of theoretical understanding into practical systems. Presents novel methodologies for emotion recognition and validation, including the development of universal testing protocols and emotion-aware interfaces.

Chapter 4: Sensing Technology and Calibration
Addresses the practical challenges of low-impact unobtrusive sensing solutions. Focuses on novel approaches to sensor calibration and alternative measurement techniques that minimize user intervention while maintaining reliability.

Chapter 5: Conclusions and Future Work
Synthesizes the findings from previous chapters and discusses their implications for the field. Identifies emerging opportunities and challenges in developing unobtrusive technological solutions.

Appendix A: Interfaces presented in the ASSISI experiment
Shows the form and pretest presented to subjects during the ASSISI experiment, presented in Chapter 2.

Appendix B: Explorations in Art and Perception
Presents complementary research on human perception through artistic ex-

pression, including studies on artificial creativity and mathematical foundations of aesthetic perception.

Chapter 2

Crossmodal Perception and Emotion

Wassily Kandinsky, "Yellow-Red-Blue" (1925), oil on canvas, 127 × 200 cm

2.1 Theoretical Background

2.1.1 Introduction to Crossmodal Perception

In many everyday situations, our senses are bombarded by different unisensory signals. To gain the most accurate and least variable estimate of environmental stimuli and properties, our brain needs to combine individual noisy unisensory perceptual estimates that refer to the same object, while keeping those estimates belonging to different objects or events separate [15]. This process of combining information from different sensory modalities has been historically viewed within a modular paradigm, with vision considered the dominant sensory modality, self-contained and independent of other senses [15]. However, research over the past decades has revealed that visual perception can be strongly altered by sound and touch, and such alterations can occur even at early stages of processing, as early as primary visual cortex [15, 16, 17, 18]. This view of vision as the dominant modality has been reinforced by classic studies of crossmodal interactions, where experimenters artificially imposed a conflict between visual information and information conveyed through another modality, and reported that the overall percept is strongly dominated by vision [15, 19, 20]. For example, in the ventriloquism effect, the perceived location of sound is captured by the location of the visual stimulus [15, 19, 20]. Visual capture of location also occurs in relation to proprioceptive and tactile modalities [21, 22]. These effects are quite strong and have been taken as evidence of visual dominance in perception [15]. Even for a function that is generally considered to be an auditory function, namely speech perception, vision has been shown to strongly alter the quality of the auditory percept, as demonstrated by the McGurk effect where pairing the sound of syllable /ba/ with the video of lips articulating syllable /ga/ induces the percept of syllable /da/ [23, 15]. The ability to integrate information from multiple senses provides several advantages in our interaction with the environment. Each sense is optimal under different circumstances, and collectively they

increase the likelihood of detecting and identifying events or objects of interest. However, these advantages are surpassed by those afforded by the ability to combine different sources of information. In this case the integrated product reveals more about the nature of the external event and does so faster and better than would be predicted from the sum of its individual contributors [15]. A striking example of this integration is shown in Figure 2.1, where the perception of visual motion can be dramatically altered by the presence of sound, demonstrating how our brain integrates information across modalities to create a coherent perceptual experience [24, 25]. To achieve successful integration, the brain must solve what is known as the "binding problem" - determining which of the many stimuli presented in different modalities at any one time should be bound together [26, 27]. This is a complex computational challenge, as the brain must determine not only which signals belong together but also how to optimally combine them given their different reliabilities and the potential conflicts between them. Recent theoretical frameworks suggest that the brain may solve this problem through Bayesian inference, weighing different sources of information according to their reliability and prior knowledge about their relationship [26, 28, 29, 30].

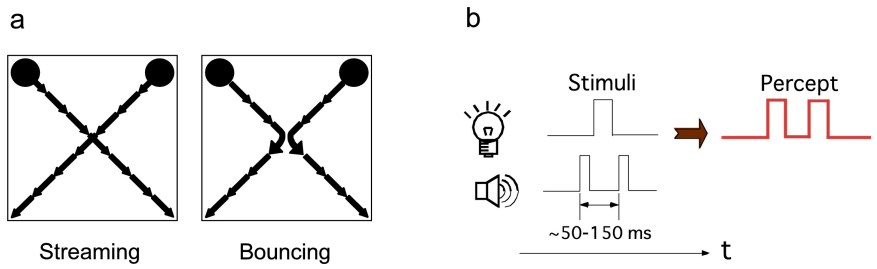


Figure 2.1. Examples of crossmodal perceptual effects [15]. (a) The streaming-bouncing illusion: Two identical visual objects moving towards each other can be perceived either as streaming through or bouncing off each other. The addition of a sound at the moment of coincidence biases perception towards bouncing [24]. (b) Temporal integration window: Visual and auditory stimuli presented within approximately 50-150ms can be integrated into a unified percept [15, 31, 32].

Research has identified several key factors that influence this binding process. Traditionally, the majority of cognitive neuroscience research on multisensory perception has focused on understanding the spatial and temporal factors modulating multisensory integration [26, 33, 22]. Broadly speaking, multisensory integration is more likely to occur the closer the stimuli in different modalities are presented in time [26, 31, 32]. The temporal window of integration, typically ranging from about 50 to 150 milliseconds (Figure 2.1b), provides flexibility in combining signals that may have different processing latencies while still maintaining their causal relationship [26]. Spatial coincidence has also been shown to facilitate multisensory integration under some conditions, although this is not universal [15, 34, 35]. Beyond these basic spatial and temporal factors, there has been increasing interest in the role of both semantic congruency and crossmodal correspondences in constraining the binding problem [26]. Semantic congruency refers to the matching of stimulus identity or meaning across modalities, while crossmodal correspondences describe systematic associations between different sensory features, such as between auditory pitch and visual size or brightness [36, 37]. These additional factors may provide yet another important means of determining which stimuli should be integrated, alongside the more traditionally studied spatial and temporal coincidence [26]. Of particular interest is how these various factors interact with emotional processing, as emotional content may serve as an additional binding factor in multisensory integration [38, 39]. Jessen and Kotz (2013) have suggested that emotional visual information may allow more reliable predicting of auditory information compared to non-emotional visual information, and Palmer et al. (2013) found that the emotional associations of music and colour were strongly correlated [38, 39].

2.1.2 Fundamentals of Multisensory Integration

At the core of crossmodal perception is the ability of the brain to integrate information from multiple sensory modalities. This process of mul-

tisensory integration has been the focus of extensive research in cognitive neuroscience. On a basic level, multisensory integration involves the combination of individual unisensory perceptual estimates that refer to the same object or event, while keeping separate those estimates belonging to different objects [15, 26]. This is a crucial function, as our senses are constantly bombarded by a variety of signals from different modalities. By integrating relevant sensory cues, the brain can form a more accurate and stable representation of the external environment [15]. The mechanisms underlying multisensory integration involve both low-level sensory processing and higher-level cognitive factors. At the sensory level, crossmodal interactions can occur as early as primary sensory cortices, challenging the traditional view of strictly segregated sensory pathways [15, 18]. Neuroimaging and neurophysiological studies have demonstrated that stimulation in one modality can modulate neural activity in ostensibly unisensory regions of the cortex [16, 17]. A key aspect of multisensory integration is the "binding problem" - how the brain determines which of the many sensory signals presented simultaneously should be grouped together as belonging to the same object or event [26, 27]. This is a computationally complex challenge, as the brain must not only identify which signals belong together, but also how to optimally combine them given their varying reliabilities and potential conflicts. Recent theoretical frameworks, such as Bayesian integration models, propose that the brain may solve this binding problem by weighting different sensory cues according to their reliability and prior knowledge about their statistical relationships [26, 28, 29]. These models suggest that crossmodal correspondences, or systematic associations between sensory features, may provide an important constraint on the crossmodal binding process [26]. Beyond these basic principles, research has also explored the various types of crossmodal interactions that can occur, ranging from low-level sensory correspondences (e.g. between pitch and brightness) to higher-level semantic and emotional associations [36, 37]. The neural substrates underlying these different forms of crossmodal processing are an active area of investigation [18, 26]. Overall, the study of multisensory in-

tegration has revealed the dynamic and interconnected nature of sensory processing in the brain, moving away from traditional modular views. Understanding the fundamental mechanisms of crossmodal perception is crucial for advancing our knowledge of human perception and cognition, as well as informing practical applications in fields such as human-computer interaction and neurotechnology.

2.1.3 Main Theories and Models

The study of crossmodal perception has given rise to several influential theoretical frameworks and computational models aimed at explaining the mechanisms underlying multisensory integration. One prominent theory is multisensory integration theory, which posits that the brain optimally combines information from different sensory modalities in order to form the most reliable and accurate perceptual estimate of the external environment [28, 40]. This theory is grounded in the idea that sensory signals are inherently noisy, and by integrating complementary cues, the brain can reduce overall perceptual uncertainty. Computational models based on this framework, such as the influential Bayesian causal inference model, suggest that the brain may accomplish this by weighting different sensory inputs according to their relative reliability, as determined by factors like signal reliability and prior knowledge [26, 30, 29]. Another important theoretical perspective is that of crossmodal plasticity, which examines how the loss or reorganisation of input from one sensory modality can lead to compensatory changes in the processing of other modalities [41, 42]. For example, studies have shown that early visual deprivation can enhance auditory and tactile perception, as the brain reallocates neural resources to process these remaining sensory inputs more efficiently [43, 44]. This suggests a high degree of flexibility in the brain's sensory systems and their ability to adapt to changes in the sensory environment. Statistical learning approaches provide another theoretical perspective, emphasizing how crossmodal associations may arise through the internalization of statistical regularities present

in the environment [45, 27]. By tracking the co-occurrence of sensory signals, the brain can learn to bind together those that tend to reliably originate from the same source, even if they are initially unrelated. This framework helps explain the emergence of crossmodal correspondences, such as the association between auditory pitch and visual size [26, 46]. Taken together, these theoretical models offer complementary explanations for the diverse phenomena observed in crossmodal perception. While they differ in their specific mechanisms, they share the common view that the brain's ability to integrate information across senses is a fundamental aspect of human perception and cognition, with important implications for understanding sensory processing, learning, and adaptation.

2.1.4 Audiovisual Interactions

A specific focus of crossmodal perception research has been on the interplay between auditory and visual information processing. Numerous studies have demonstrated robust interactions between these two sensory modalities, highlighting their tight integration in human perception. Key studies in this area have revealed that the presentation of auditory and visual stimuli can significantly alter the perceptual processing of either modality alone. For example, Sekuler et al. (1997) showed that the addition of a brief sound at the moment of visual collision can bias the perception of two moving objects towards a bouncing, rather than streaming, trajectory [24]. Similarly, Shams et al. (2000) found that a single flash of light paired with multiple auditory beeps can induce the illusory perception of multiple flashes [25]. These crossmodal illusions suggest that the brain integrates auditory and visual information in a robust and automatic fashion. The temporal and spatial relationships between auditory and visual stimuli have been identified as crucial factors modulating these audiovisual interactions. Studies have consistently shown that multisensory integration is most likely to occur when the stimuli are presented in close temporal proximity, typically within a window of 50-150 milliseconds [15, 31, 32]. Spatial coinci-

dence has also been found to facilitate audiovisual integration under some conditions, although this is not a universal requirement [15, 34, 35]. Beyond these basic spatiotemporal factors, research has also highlighted the importance of crossmodal correspondences in audiovisual processing [26]. Systematic associations between auditory features, such as pitch, and visual features, such as size or brightness, have been shown to significantly influence how auditory and visual information is integrated and perceived [38, 39].

2.1.5 Emotion in Crossmodal Processing

The role of emotion has emerged as an important factor in understanding crossmodal perception and integration. Recent studies have suggested that emotional content can serve as an additional binding mechanism, influencing how information from different sensory modalities is combined. Jessen and Kotz (2013) proposed that emotional visual information may allow more reliable prediction of accompanying auditory information, compared to non-emotional visual cues [38]. Their research indicates that emotional associations between audiovisual stimuli can facilitate neural processing and integration of the multisensory percept. Similarly, Palmer et al. (2013) found that the emotional associations of music and colour were strongly correlated, with faster, major-mode music being paired with more saturated, lighter, and yellower colours, while slower, minor-mode music was associated with more desaturated, darker, and bluer colours [39]. These findings suggest that the emotional content of sensory signals plays a crucial role in shaping crossmodal correspondences and multisensory integration. The neural mechanisms underlying the influence of emotion on crossmodal processing are an active area of investigation. Studies have begun to explore how emotional processing interacts with the brain's systems for binding information across sensory modalities, and how this interaction may impact perceptual experience and behaviour. Overall, the research on emotion in crossmodal perception indicates that affective factors cannot be

overlooked when seeking to fully understand the complex and dynamic nature of human multisensory integration. Integrating emotional influences into theoretical models of crossmodal processing represents an important frontier for future research in this field.

2.2 The Role of Emotion in Audiovisual Associations

The investigation of audiovisual associations has a rich history dating back to the 1920s and 1930s. Early pioneers like [47] explored the relationship between sound and colour, demonstrating that non-synesthetic individuals could consistently match certain musical notes and harmonies with specific colours. This foundational work sparked decades of research into the nature of crossmodal correspondences and their underlying mechanisms. In a seminal study, [48] demonstrated that people without synaesthesia could reliably match colours with musical selections, suggesting that these associations were not limited to synesthetes but represented a more universal aspect of human perception. Building on this work, [49] made a crucial discovery: participants tended to choose similar colours for musical selections that evoked similar emotional responses. When participants disagreed about the emotional evaluation of a piece of music, their colour associations became inconsistent, providing early evidence for emotion as a mediating factor in audiovisual associations. The role of emotion as a mediating factor in crossmodal associations has been substantiated by numerous studies over the subsequent decades. Research has consistently shown that positive emotions tend to be associated with brighter, more saturated colours, while negative emotions are linked to darker, less saturated colours [50, 51]. Similarly, in the auditory domain, major musical modes are typically associated with positive emotions, while minor modes evoke more negative emotional responses [52]. The emotional mediation hypothesis suggests that crossmodal associations occur through shared emotional qualities between different sensory modalities. For instance, both a bright yellow colour and an upbeat major-mode melody might evoke feelings of happiness, leading to their association even though they are processed through different sensory channels. This hypothesis has received strong support from contemporary research showing consistent patterns of association between musical

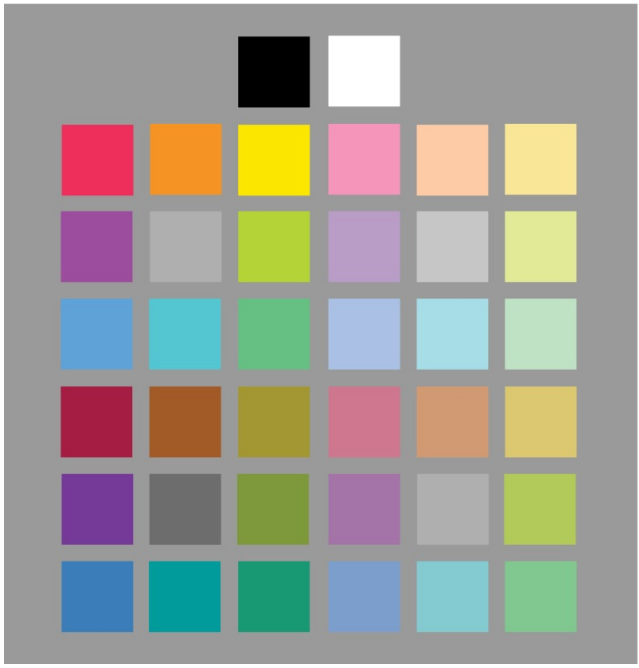


Figure 2.2. The 37 colours used in Palmer et al.'s study of music-colour associations, systematically varying in hue, saturation, and brightness. These colours were used to demonstrate systematic relationships between musical emotions and colour choices [39].

features and specific colour properties mediated by emotional responses. A significant advancement in understanding these associations came from studies examining the role of emotional congruence in audiovisual integration. [53] demonstrated that semantically congruent multisensory stimuli result in enhanced behavioural performance, while incongruent stimuli can lead to performance decrements. This finding suggests that emotional congruence plays a crucial role in how our brain processes and integrates information from different sensory modalities. The neural basis of emotionally mediated audiovisual associations has been increasingly revealed through modern neuroimaging techniques. Studies have shown that both early sensory areas and higher-order association cortices are involved in processing emotionally congruent audiovisual information [26]. The supe-

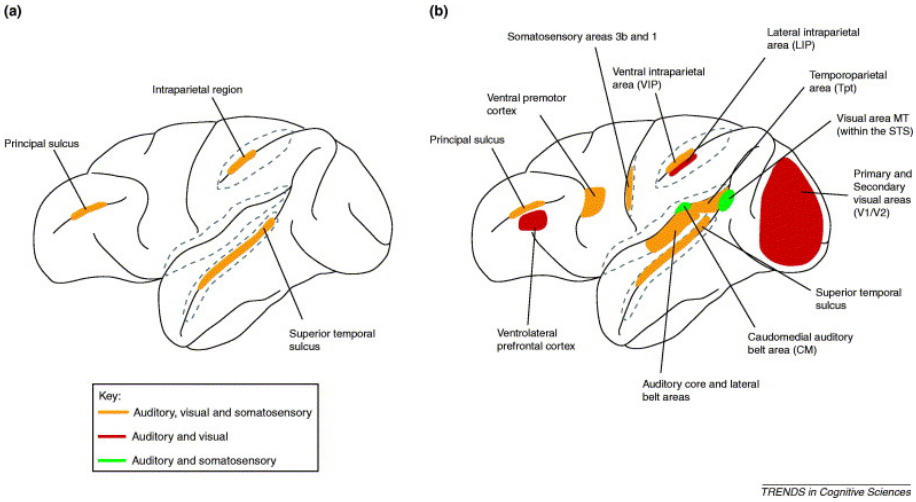


Figure 2.3. Brain regions involved in multisensory integration, showing the network of areas that process emotionally congruent audiovisual information. Adapted from [18].

rior temporal sulcus (STS) has emerged as a particularly important region for integrating emotional information from different modalities. Research has demonstrated that the STS shows enhanced activation when processing emotionally congruent audiovisual stimuli compared to incongruent combinations [18]. Recent studies have also highlighted the rapid nature of emotional audiovisual integration. Event-related potential (ERP) studies have shown that emotional congruence between auditory and visual stimuli can influence neural processing as early as 100 milliseconds after stimulus onset [54]. This suggests that emotional information plays a fundamental role in the early stages of sensory integration, rather than being a later cognitive process. The temporal dynamics of audiovisual integration are particularly relevant when considering musical stimuli. [55] found that the emotional intention of musical performers influenced listeners' colour choices, with specific correlations between expressed emotions and preferences for hue, saturation, and brightness. These associations were found to be remarkably consistent across different musical genres and instrumental timbres. [39]'s comprehensive study provided strong evidence for emotion as the critical

mediating factor in music-colour associations. Using a systematic approach with carefully controlled stimuli (see Figure 1), they demonstrated that people consistently chose colours whose emotional associations matched the emotional qualities of the music they were hearing. This matching occurred across cultures, suggesting a universal basis for these emotionally mediated associations. Furthermore, research has shown that these audiovisual associations are not static but can be influenced by context and experience [56]. Studies investigating crossmodal correspondences in children and adults have revealed both developmental constants and changes in how emotions mediate sensory associations. This suggests that while some aspects of emotional mediation in crossmodal associations may be innate, others are shaped by cultural and personal experience. The understanding of emotionally mediated audiovisual associations has important implications for various fields, including multimedia design, artistic expression, and therapeutic applications [54]. For instance, this knowledge can inform the development of more effective audiovisual interfaces, enhance the emotional impact of multimedia artworks, and contribute to the design of sensory-based therapeutic interventions. These findings collectively support a model where emotion serves as a fundamental binding factor in crossmodal perception, facilitating the integration of information across sensory modalities. This integration appears to operate at multiple levels, from basic sensory processing to higher-order cognitive evaluation, suggesting that emotion plays a crucial role in how we construct our unified sensory experience of the world [26].

2.3 Research Objectives and Chapter Overview

This chapter investigates the foundational role of emotions in crossmodal perception through two interconnected studies that explore how emotional responses mediate audiovisual associations. The research addresses fundamental questions about the nature of crossmodal correspondences and their emotional underpinnings whilst providing practical insights for the development of emotionally congruent multimodal interfaces.

The first study, presented during the 28th International Conference on Auditory Display (ICAD) in Norrköping, Sweden in 2023 [57], examines colour-sound mapping preferences across different age groups, investigating how children and adults associate musical elements with visual properties. This work aims to understand developmental aspects of crossmodal associations and establish whether consistent mapping patterns emerge across age groups. The investigation employs a novel experimental protocol designed to capture both explicit preferences and implicit associations, providing insights into the development of crossmodal processing.

The second study presents a comprehensive quantitative analysis of emotion-mediated audiovisual associations. Through rigorous experimental validation, this research examines how emotional responses influence the relationship between visual stimuli and musical characteristics. The study introduces novel methodologies for measuring emotional mediation in crossmodal perception and develops predictive models for audiovisual mappings based on emotional content.

Together, these investigations advance our understanding of emotion's role in crossmodal perception whilst establishing methodological frameworks for future research in emotionally-aware interface design. The findings contribute to both theoretical knowledge of crossmodal processing and practical applications in human-computer interaction, particularly in contexts where maintaining emotional congruence across modalities is crucial.

2.4 Investigating Colour-Sound Mapping in Children and Adults

Foundational Concepts and Research Aims

Sonification, the use of non-speech audio to convey information or perceptualise data [58], has traditionally relied on individual sounds or tones to represent visual stimuli. However, given the role of emotion in crossmodal associations, as evidenced in the preceding paragraphs, we propose an alternative approach. Our study investigates audiovisual associations using musical chords based on Western harmonic principles [59]. By systematically varying the consonance and dissonance of these chords, we aim to explore how complex musical stimuli influence crossmodal perception, extending existing research in this domain. Furthermore, we compare the audiovisual mappings of children and adults to examine developmental differences. Previous studies suggest that some audiovisual associations may be innate [60], while others are learned through cultural exposure [61]. By including both age groups, we seek to untangle the innate and acquired aspects of these associations. This study contributes to our understanding of how musical harmony shapes crossmodal correspondences and highlights the developmental trajectory of audiovisual perception. The findings provide a foundation for future research exploring the emotional dimensions of these phenomena.

Association criterion design

The first step in designing the association was to select the colour model and the type of musical stimulus. To do so, we considered the literature on synesthetes who experience coloured hearing, where the sight of colour automatically leads to the involuntary experience of sound, and studies on crossmodal correspondences between visual and acoustic dimensions.

Colour model After reviewing different studies on the colour-sound association, we chose to adopt the Hue, Saturation, Brightness (HSB) colour model: Hue refers to the basic colour tone, Saturation defines the intensity or purity of a colour, with higher saturation indicating a more vivid colour, and Brightness refers to the perceived brightness of a colour, with higher brightness indicating a lighter colour. According to a recent review on coloured hearing [62], these three components are very commonly associated with the correspondence between colour and sound:

1. Hue is often associated with timbre [63, 64] and historically linked to pitch [65, 66].
2. Saturation has been reported to be associated with loudness and pitch [67].
3. Brightness is often seen in relation to pitch and loudness [68, 69].

Musical model Since the objective of this work is to design an association criterion that allows users to feel the emotion they would upon seeing a colour while listening to music, it was crucial to choose a musical model that could elicit a sentiment. For this reason, we decided to use an auditory stimulus that reproduces sounds of common instruments with notes and harmony in accordance with the rules of western music.

The key characteristics were two: it had to be representable on a scale with sufficiently high resolution so that each colour could be linked to a single sound, and it had to remain unchanged over time regardless of the surroundings, as sensory substitution devices are to be taken everywhere. From the various mappings that were reported above, the obvious choice for the music model would be to use a representation including pitch, loudness, and timbre. Out of these, pitch is the only one that satisfies the aforementioned conditions: timbre can be difficult to discern and is not easily representable in a scale, and loudness is heavily influenced by outside noises.

For all these reasons, the selected musical stimulus was a chord, which was also one of the forms of Scriabin's synesthesia [70]. Chords were gen-

erated according to occidental classical music harmony. The following parameters of the chord were selected:

1. The root tone (RT), which is one out of the 12 notes in the chromatic scale (i.e. C, C#/D, E, ...).
2. The octave on the piano (OC), which is one out of 7 (i.e. C1, C2, ...).
3. The mode of the chord (MO), which is one out of up to 12 (i.e. minor, major, ...).

The chord is played with all the notes at the same time and can be heard with a frequency of one chord per second.

Correspondence between stimuli

Once the models for colour and music were chosen, one out of the 6 possible mappings of the musical-colour parameter associations (see Table 2.1) had to be chosen. Since literature on the topic does not provide a clear indication, the best course of action was to gather data from subjects in a trial. In this way, the outcome of the study would be an association logic that elicits the same emotions for most of the population so that there is the greatest chance that devices developed using such logic will not be rejected for lack of emotional stimulus.

When citing the possible mappings later in this paper, they will be called with the three musical parameters respectively associated with Hue, Saturation, and Brightness (for example, the mapping Hue-OC, Saturation-RT, Brightness-MO will be called OC-RT-MO).

Tailoring of the logic

In order to select which musical parameter had to be associated with each color parameter, we designed an experimental trial. This study received the approval of the university's ethical committee on February 16, 2022, with the clinical studies register number 2021.236.

Hue	Saturation	Brightness
OC	MO	RT
OC	RT	MO
MO	OC	RT
RT	OC	MO
MO	RT	OC
RT	MO	OC

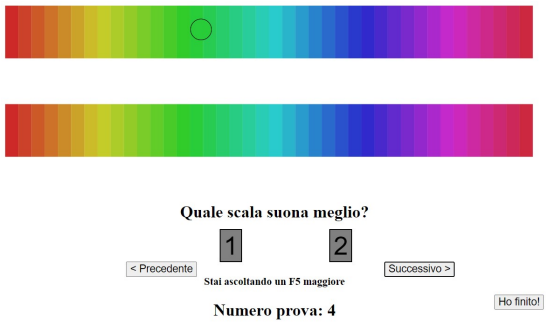
Table 2.1. Possible mappings between HSB and musical parameters

Protocol The test was administered remotely via a web app. The graphical interface was developed using `p5.js`[1], the audio was generated with `soundfont-player`[2], and the backend was developed using `node.js`[3].

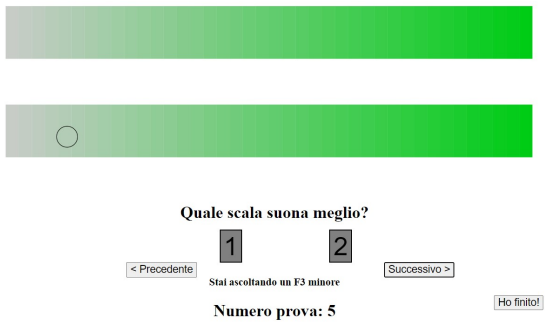
Subjects were shown two color spectra (see Figure 2.4) and could move the cursor on them to hear a chord played by a piano repeatedly. The two scales had different color-music mappings, so each color played differently in the two scales. They were asked to set the audio and screen settings to their preferences and to choose which of the two scales sounded more pleasant. They repeated the selection three times for each of the 6 possible mappings, with two of the values of the HSB model set and only the last one changed in the scale. For example, saturation and brightness were set to an intermediate value and they were shown a scale of hue. The total number of selections to be completed was 45, and they were presented to all subjects in the same order. the test lasted an average of 40 minutes, with a minimum time of 30 minutes and a maximum of 60 minutes.

Each scale selection was saved in the results, which could be downloaded as a JSON file at the end of the test and sent to the author. If at any point they felt tired, they could click on the "I'm done!" button and save their results. They could then resume the test at a later time.

[1] `https://p5js.org/`
[2] `https://www.npmjs.com/package/soundfont-player`
[3] `https://nodejs.org/`



(a) Shades of hue



(b) Shades of saturation

(c) Shades of brightness

Figure 2.4. Graphical Interface of the test: The two scales play different chords for the same colour. Participants have to choose which one sounds better to them, they can go back to the previous selection and move on to the next one. They are shown the name of the musical chord they are hearing and the number of the selection they are currently at and they can click the "Finish!" button if they are tired. As all the participants were from Italy, the test was administered in Italian.

Participants Two groups of participants took part in the study: a group of 8 children aged between 7 and 11 years old (with a mean age of 9.5 years, 37.5% were female) and a group of 8 adults aged between 25 and 36 years old (with a mean age of 28 years, 50% were female). A summary of participant demographics is provided in Table 2.2

(a) Children Group

Subject	Sex	Age
1	M	11
2	M	9
3	F	8
4	M	9
5	F	11
6	F	11
7	M	7
8	M	8

(b) Adults Group

Subject	Sex	Age
1	F	29
2	F	28
3	F	25
4	M	26
5	M	27
6	M	25
7	M	36
8	F	28

Table 2.2. Demographic details of study participants [57]

Before starting the test, all participants signed the privacy policy and informed consent form.

Data Analysis

Data collection For each participant, we extracted two pieces of information:

- The number of times each mapping was chosen: each pairing was presented to the subjects 5 times for comparison with the other possible mapping. For each couple of mappings, they had to choose which of the two mappings they preferred three times, one for hue, saturation, and brightness. For each comparison between mappings, we considered the mapping selected if it was chosen at least 2 out of 3 times for the color parameters comparison.

- The number of times each color-musical parameter pairing was chosen: each pairing could be selected a maximum of 5 times.

Once we had extracted these values from each subject, we computed the mean and standard error of the results for each group and for all subjects to determine the most voted mappings. We also applied the ANOVA test to the mappings for both groups and all subjects to verify the significance of at least one preferred mapping. Before performing the test, the Lilliefors test was applied to the mapping results to verify their Gaussian distribution.

Results The Lilliefors did not have positive results for the children's, adults', and combined groups, with p-values of 0.0026, 0.0106, and 0.001, respectively. Results from the ANOVA are shown in Table 2.3

Population	df	F	p-value
Children	5	2.31	0.06
Adults	5	1.76	0.14
Children and adults	5	1.68	0.15

Table 2.3. ANOVA test results on mappings: p-values less than 0.05 would indicate rejection of the null hypothesis of equal means at 5% significance level [57]

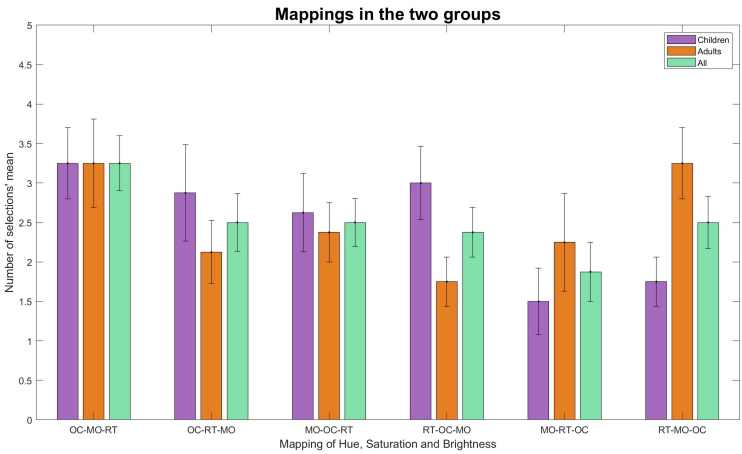
The ANOVA results for children suggest a trend towards a difference in the ratings of the mappings, with a value close to 0.05. In contrast, the differences in adults' choice of mappings are not significant. For this reason, when combining the results of children with those of adults, the significance of the children's group is lost.

The results of the data collection are shown in Table 2.4.

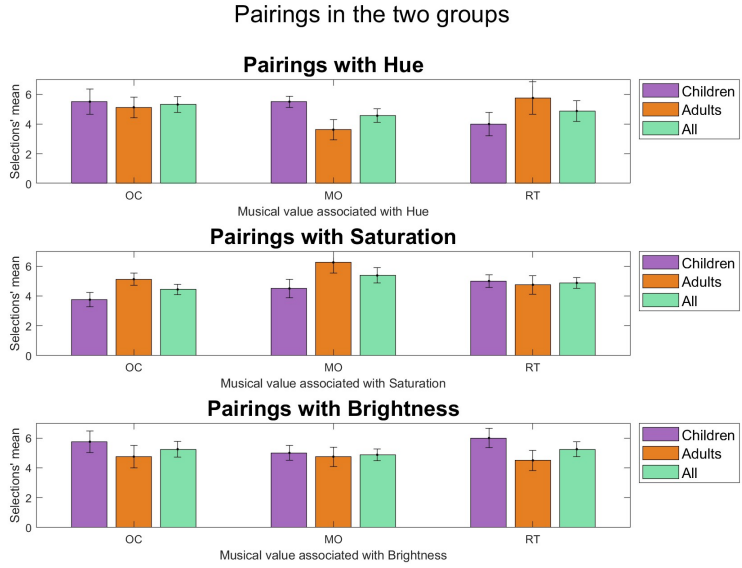
The differences between the two groups can be seen in Figure 2.5.

Children's mappings' results can be described as follows:

- A strong preference for the OC-MO-RT mapping (mean value = 3.25) with a low variance (0.45), indicating compactness in the group.



(a) Differences in mappings



(b) Differences in pairings

Figure 2.5. Comparison of mapping and pairing preferences between children and adults [57]

Table 2.4. Results of mapping preferences and pairing ratings [57]

(a) Mapping ratings in children

Hue	Saturation	Brightness	Mean
OC	MO	RT	3.3 ± 0.45
OC	RT	MO	2.9 ± 0.61
MO	OC	RT	2.6 ± 0.50
RT	OC	MO	3.0 ± 0.46
MO	RT	OC	1.5 ± 0.42
RT	MO	OC	1.8 ± 0.31

(b) Mapping ratings in adults

Hue	Saturation	Brightness	Mean
OC	MO	RT	3.3 ± 0.56
OC	RT	MO	2.1 ± 0.40
MO	OC	RT	2.4 ± 0.45
RT	OC	MO	1.75 ± 0.40
MO	RT	OC	2.25 ± 0.62
RT	MO	OC	3.25 ± 0.38

(c) Pairings ratings in children

	OC	MO	RT
Hue	5.5 ± 0.79	3.8 ± 0.46	5.8 ± 0.68
Saturation	5.5 ± 0.35	4.5 ± 0.59	5.0 ± 0.47
Brightness	4.0 ± 0.73	5.0 ± 0.40	6.0 ± 0.61

(d) Pairings ratings in adults

	OC	MO	RT
Hue	5.1 ± 0.69	5.1 ± 0.40	4.8 ± 0.75
Saturation	3.6 ± 0.68	6.3 ± 0.73	4.8 ± 0.65
Brightness	5.8 ± 1.08	4.8 ± 0.62	4.5 ± 0.68

(e) Mappings ratings in subjects

Hue	Saturation	Brightness	Mean
OC	MO	RT	3.3 ± 0.47
OC	RT	MO	2.5 ± 0.53
MO	OC	RT	2.5 ± 0.70
RT	OC	MO	2.4 ± 0.44
MO	RT	OC	1.9 ± 0.64
RT	MO	OC	2.5 ± 0.52

(f) Pairings ratings in subjects

	OC	MO	RT
Hue	5.3 ± 0.79	4.4 ± 0.48	5.3 ± 0.76
Saturation	4.6 ± 0.60	5.4 ± 0.70	4.9 ± 0.58
Brightness	4.8 ± 0.93	4.9 ± 0.48	5.2 ± 0.73

- A slightly less preference for the RT-OC-MO mapping (mean value = 3.00) with a low standard error (0.41).
- A weak preference of the MO-RT-OC mapping (mean value = 1.5) with a low variance (0.42).
- An equally weak preference of the RT-MO-OC mapping (mean value = 1.8) with a very low standard error (0.31).

Their results on pairings are:

- The best sonification of the Hue is through the Root Tone (mean value = 5.8), followed by the Octave (5.5). Both these values, however,

have a slightly high value of standard error (i.e., 0.68 and 0.79 respectively), suggesting lower agreement in the group. On the other hand, the pairing Hue-Mode is the least voted with high agreement.

- The best pairing for saturation is the Octave, which collected a high number of preferences with a low standard error (mean value = 5.5, standard error = 0.35), followed by the Root Tone (mean value = 5.0, standard error = 0.47).
- Lastly, the most voted pairings for brightness were the Root Tone (mean value = 6.0), whose standard error is a little high (0.61). The Mode also received a high score (mean value = 5.0) with high agreement.

As for the adults, their preferences on mappings are described below:

- The preferred mapping is OC-MO-RT (mean value = 3.3, standard error = 0.56).
- RT-MO-OC received a similar score (3.25), but with a lower standard error (0.38), indicating a higher agreement in the group.
- They show a weaker preference for the mapping RT-OC-MO (mean value = 1.75) with a low standard error (0.40).
- Similarly, they voted the mapping OC-RT-MO the least with a high agreement (mean value = 2.1, standard error = 0.40).

Their choices on the pairings are listed below:

- For the sonification of Hue, they like Octave and Mode equally (5.1), with the first one having a higher standard error, indicating the more agreement in the second choice.
- All the values of the saturation have fairly high standard error (ranging from 0.68 to 0.65), suggesting a general low agreement in this parameter. Despite this, a strong preference for the mode (mean value

= 6.3) and a weak preference of the Octave (mean value = 3.6) can be seen.

- The higher pairing for brightness, which is the Octave (mean value = 5.8) have a very high standard error (1.08), indicating a high disagreement in the group.

When analysing the two groups as one, the following results emerged:

- There is a preference for the mapping OC-MO-RT (mean value = 3.3) with a relatively low standard error (i.e. 0.47) indicating agreement in the whole group.
- There is a weak preference of the mapping MO-RT-OC (mean value = 1.9).
- The pairings value belong in a very small range (from 4.6 to 5.4) and not much can be said about preferences.

Discussion

Results from the two groups show some differences, which can be interpreted as an indication of a different perception of color and music. In particular, there is a high discrepancy in their preferences in two mappings:

- RT-MO-OC, which was the second most voted by adults but the least voted by children.
- RT-OC-MO, which was the second most voted by children but the least voted by adults.

These differences can be explained by the pairing with saturation: children prefer the pairing with the Octave the most, while it is the worst for adults, and adults prefer the pairing with the Mode the most, while it is the worst for children.

Despite this, both groups indicate with high agreement that the mapping OC-MO-RT is the preferred one. This can be explained by the strong

preference for the pairing Hue-OC in both groups, for Brightness-RT in children, and for Saturation-MO in adults.

This result suggests that, despite their differences, children and adults agree on this type of mapping, which could be the best one for colour sonification.

It is important to note that, despite the agreement, the ANOVA test revealed that the results were close to be significant for the children's group, whereas the results of the adults and the entire group were not significant. Therefore, the preference for mapping can only be expected in this particular group.

The limited number of subjects considered must be taken into account when drawing conclusions, which can only be preliminary at this point and should be confirmed by a higher number of data.

Conclusion and future developments

In this paper, the authors presented the design of an association criterion between colour and music with the objective of finding the most natural to all subjects.

Three musical parameters and the HSB colour representation were chosen for this task, and an experimental protocol was developed to find the best mapping. The protocol was applied to two groups of subjects: the first consisted of 8 children, and the second one of 8 adults. Their preferences on mappings and pairings were registered and confronted.

Results show that children and adults present several differences in their preferences, but both like the mapping Hue-Octave, Saturation-Mode, Brightness-Root Tone the best. This preference is a result of the strong liking of the pairing Hue-Octave by both groups, of the Saturation-Mode by the adults, and of the Brightness-Root Tone by the children.

Their agreement on the same mapping is promising for the identification of a new mapping that will enable a more natural sonification, but it is yet a preliminary result that will have to be confirmed by a higher number of data collected from more subjects.

These results contribute to our understanding of how complex auditory stimuli, such as chords, can influence audiovisual perception and highlight the role of development in shaping these associations. The study provides a foundation for future research exploring the impact of musical harmony on crossmodal correspondences and the emotional dimensions of audiovisual perception.

2.5 Experimental Validation and Quantitative Analysis of Emotion-Mediated Audiovisual Crossmodal Associations

2.5.1 Introduction and Rationale

As detailed in Chapter 2.2, substantial research has highlighted the importance of emotion in audiovisual crossmodal associations, with emotions serving as a crucial bridge between auditory and visual perception. The aim of this experiment is to extend that understanding by quantifying the precise role of emotional responses in mediating these associations, specifically in how music-evoked emotions influence the visual characteristics that individuals create or modify, and vice versa. This research not only explores the bidirectional nature of these associations but also examines the potential for predictive models in this context.

To ensure robust and reproducible findings, we began with the design, implementation, and validation of an experimental protocol specifically aimed at capturing participants' bidirectional associations between images, music and their emotions. This validation confirmed the protocol's suitability and reliability, forming a solid foundation for our further data acquisition and analyses. This process was presented during the 2nd Advanced Course on Artificial Intelligence & Neuroscience in Certosa di Pontignano, Italy, in 2022 [71].

Following protocol validation, we conducted two studies with distinct aims:

- The first study, conditionally accepted in PLOS ONE pending minor revisions, focuses on quantifying the role of emotions in mediating crossmodal associations, analysing whether certain emotional responses (e.g., sadness, nostalgia, amazement) consistently influence visual parameters like Brightness, Saturation, and Spatial Dispersion. This analysis will reveal the strength and nuances of emotion's impact on

the crossmodality, providing a basis for predicting these associations.

- The second study, currently under review in Computers in Human Behavior: Artificial Humans, takes a focused look at the first phase of data acquisition, using predictive models to examine the bidirectional nature of these crossmodal associations. We aim to assess the capability of these models to predict visual parameters from emotional cues and, conversely, to infer emotional responses based on the visual characteristics of images generated by participants. This bidirectional modelling not only enhances our understanding of the mechanisms underlying audiovisual associations but also supports potential applications in affective computing.

By clarifying and quantifying the emotional mechanisms linking auditory and visual modalities, this research opens the way for various applications in sensory substitution and inclusive tools. For instance, it suggests pathways to improve audiovisual sensory substitution systems through advanced affective computing tools (as discussed further in Chapter 3). This work also supports the development of inclusive technologies that could enrich the experience of exhibitions, concerts, and other interactive events, making them more accessible and emotionally resonant for diverse audiences.

2.5.2 Validation of the Experimental Protocol

Initial Protocol Development and Alpha Testing

The development and validation of the experimental protocol followed an iterative approach to ensure robust and reliable data collection for investigating audiovisual associations and their emotional components. The protocol's initial design was grounded in previous research on crossmodal associations highlighted in the previous chapter, whilst incorporating novel elements to address the specific requirements of our investigation into emotion-mediated audiovisual interactions.

The alpha testing phase was conducted locally with a balanced sample of 10 Italian participants. This preliminary testing was crucial in identifying two significant limitations in the initial protocol design. To ensure optimal comprehension and eliminate potential language-related confounds, the protocol was implemented entirely in Italian, as all participants across the alpha, beta, and final testing phases were native Italian speakers. This included all instructions, interface elements, questionnaires and response labels.

The first limitation concerned the emotion classification system. The protocol initially employed Ekman's six basic emotions (happiness, sadness, fear, disgust, anger, surprise) [72], a framework widely used in emotional research. However, participant feedback consistently indicated that these basic emotions were inadequate for capturing the nuanced emotional responses evoked by musical stimuli. This finding aligned with previous research suggesting that music-induced emotions often extend beyond basic emotional categories [73]. Consequently, the protocol was modified to incorporate an adaptation of the Geneva Emotional Music Scale (GEMS) [74], a validated framework specifically designed for assessing music-evoked emotions.

The second significant finding pertained to the visual parameters of the image generation interface. The original design utilised a grey background as a presumed neutral base for participant-generated images. However, qualitative feedback revealed that this choice inadvertently introduced an emotional bias, with participants reporting that the grey background induced an inherent feeling of sadness across all generated images, regardless of other parameter settings. This unintended emotional priming was addressed by modifying the background to an off-white colour, providing a more emotionally neutral baseline for image generation.

These initial findings proved instrumental in refining the protocol's capacity to accurately capture emotional responses in audiovisual associations. The modifications implemented following the alpha testing phase significantly enhanced the protocol's sensitivity to music-induced emotions

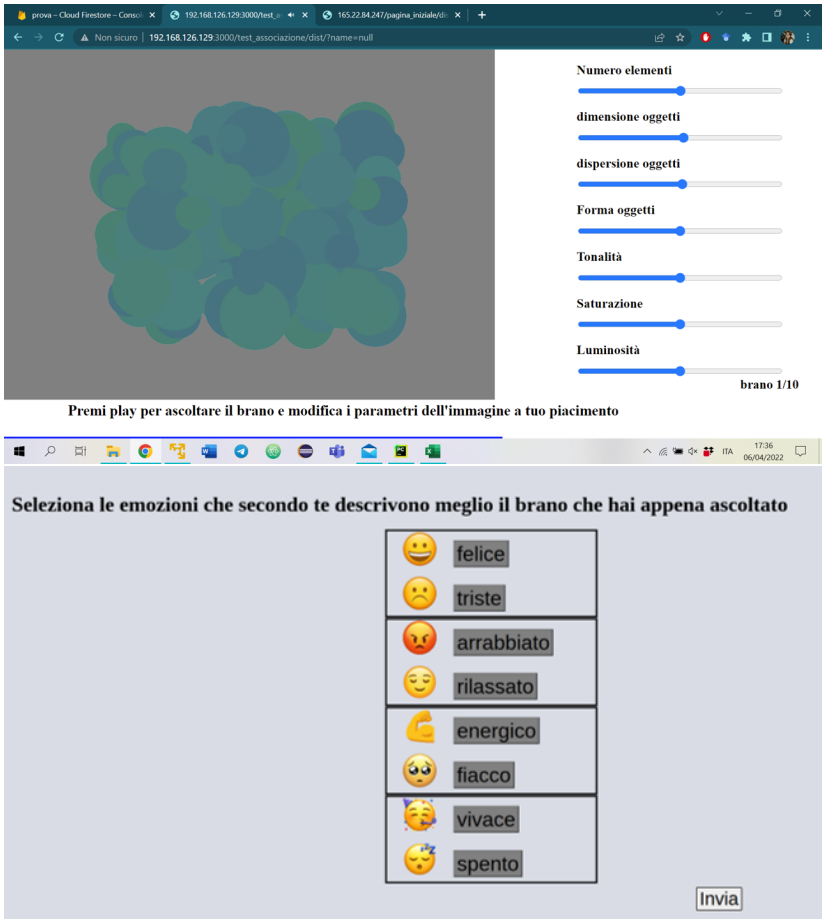


Figure 2.6. Screenshots from the alpha testing phase showing the original interface with Ekman's emotion categories and grey background

and eliminated unintended visual biases, establishing a more robust foundation for the subsequent beta testing phase.

Beta Testing Implementation

Following the refinements implemented after the alpha testing phase, a comprehensive beta testing was conducted to validate the modified protocol. This phase focused on both the technical robustness of the implementa-

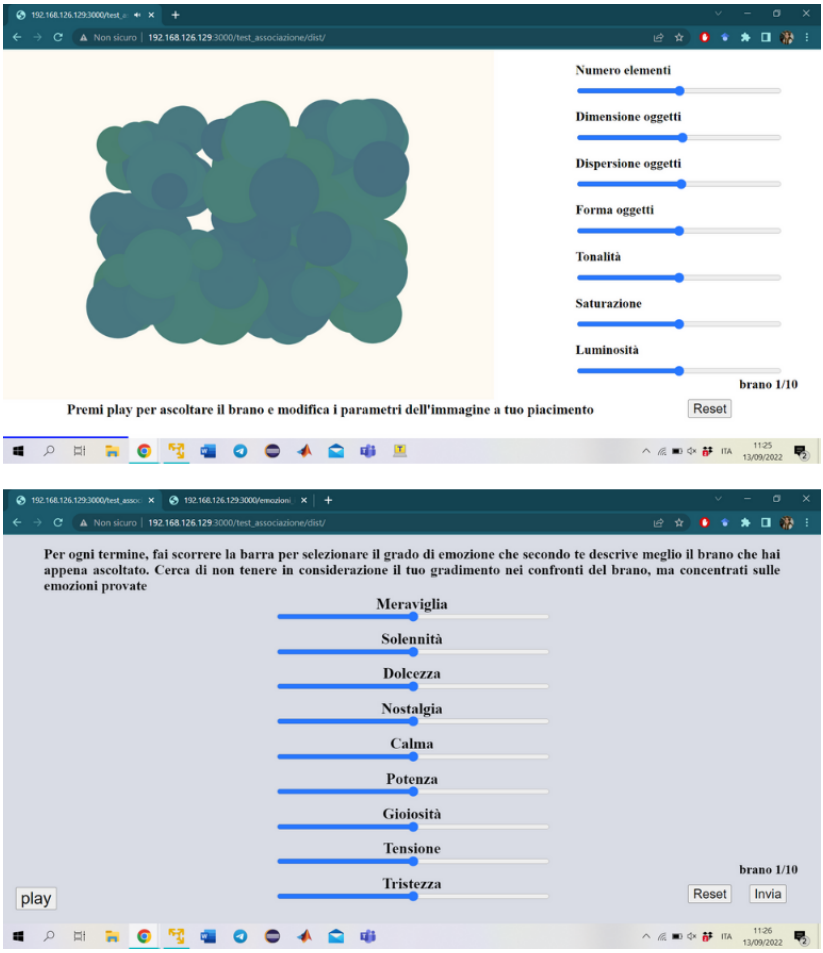


Figure 2.7. Screenshots from the alpha testing phase showing the modified interface incorporating GEMS and off-white background.

tion and the protocol's effectiveness in capturing audiovisual associations.

The experimental platform was developed as a web-based application to ensure widespread accessibility and standardised testing conditions. The technical implementation utilised a full-stack JavaScript architecture, with the front-end visualisation developed in p5.js [4] and the back-end function-

[4] https://p5js.org/

ality implemented using Node.js [5]. This architecture was chosen for its capability to handle real-time interactive visualisations and complex data collection requirements.

Data collection and storage were managed through a Firebase [6] database system, selected for its reliability in handling concurrent users and its robust data validation capabilities. The system was designed to store participants' responses and interactions systematically, including:

- Preliminary questionnaire responses
- Pre-test performance metrics
- Parameter adjustments during image generation
- Temporal data on user interactions
- Emotional response measurements

The implementation included several key technical features to ensure data quality and participant engagement:

1. **Session Management:** The system allowed participants to complete the protocol across multiple sessions, implementing secure user authentication and progress tracking to prevent fatigue-induced bias.
2. **Environmental Controls:** Specific technical requirements were enforced, including:
    - High-resolution display specifications
    - Mandatory headphone use for standardised audio presentation
    - Automatic detection and warning of night-shift or colour-modification settings
    - Browser compatibility verification

[5] `https://nodejs.org/`
[6] `https://firebase.google.com/`

3. **Quality Assurance:** The implementation incorporated:
    - Real-time data validation
    - Automatic error detection and handling
    - Response time tracking
4. **User Interface Design:** The interface was optimised for:
    - Intuitive parameter control
    - Real-time visual feedback
    - Clear instruction presentation

To ensure standardised testing conditions, participants received detailed instructions regarding environmental requirements and testing procedures. The system included automated checks for compliance with these requirements, ensuring data quality and reliability across different testing environments.

The beta implementation was tested with 14 participants, focusing on both the technical robustness of the platform and the effectiveness of the protocol modifications implemented following the alpha testing phase. This testing phase was crucial in validating the platform's capability to reliably capture complex audiovisual associations and their emotional components in a remote testing environment.

A fundamental change to the protocol was made after the beta phase, due to the participants reporting the occurrence of a familiarity bias during the test. Familiarity bias can be described as the tendency to seek confirmation of expectations, retaining, or avoiding abandoning favoured hypotheses or choices [75]. During the beta testing of the protocol, some subjects reported that they felt inclined to align their choices in the second phase with their selections in the first phase instead of answering spontaneously, potentially introducing a familiarity bias in the results. To address this concern, a decision was made to divide the subjects into two groups: Group A, which completed the entire protocol, and Group B, which only completed the pretest and the second phase.

The purpose of this division was to establish a standard reference for a typical emotional response to the images by observing Group B's results. Subsequently, the responses of Group A could be compared to this reference to determine if there were any deviations to be addressed to the completion of phase one, thereby assessing the presence of a familiarity bias.

Validation Results and Protocol Refinement

The beta testing phase, conducted with 14 participants, provided comprehensive validation of both the technical implementation and the experimental protocol. The results demonstrated the protocol's effectiveness across multiple dimensions whilst identifying areas for final refinement.

Technical and User Experience Validation The web-based platform demonstrated exceptional robustness throughout the testing phase. The system successfully collected complete datasets from all participants with no reported system errors or data loss. All participant interactions were successfully stored and retrieved from the database, whilst the platform maintained consistent performance across various browser environments. The audio playback and parameter control functionality proved particularly reliable, ensuring consistent experimental conditions across all sessions.

The protocol's usability was strongly validated through participant feedback and interaction analysis. All participants successfully completed the full protocol, with consistently positive feedback regarding the interface's accessibility and intuitive design. The carefully planned session duration proved appropriate, with no participants reporting significant fatigue effects. Furthermore, the implemented multi-session capability proved valuable, allowing participants to effectively manage their engagement with the protocol according to their individual needs and circumstances.

Data Quality Assessment Preliminary analysis of the collected data revealed meaningful patterns and correlations, supporting the protocol's validity and enabling the development and validation of analytical tools for

subsequent data collection phases:

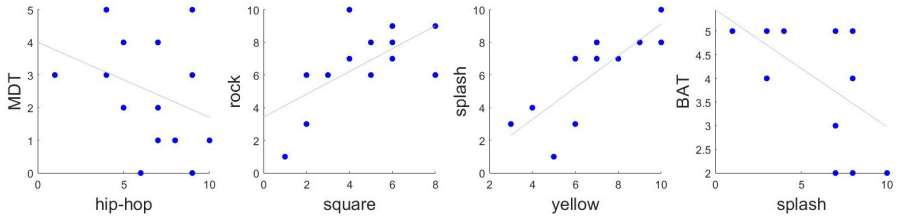


Figure 2.8. Examples of correlations identified during beta testing: (a) Relationship between musical preferences and melodic discrimination test performance, (b) Correlation between colour preferences and shape selection, (c) Association patterns between musical genres and emotional responses.

The analysis revealed several significant relationships:

- Correlation between musical genre preferences and performance in the melodic discrimination test, particularly notable in the case of Hip-Hop preferences showing an inverse relationship with test performance
- Consistent associations between colour preferences and shape selections, suggesting robust crossmodal patterns
- Systematic relationships between musical parameters and emotional responses, validating the effectiveness of the GEMS implementation

Protocol Refinements Following the beta testing phase, several refinements were implemented to optimise the protocol. The interface underwent minimal enhancements to improve user interaction. Parameter control labels were clarified to ensure intuitive understanding, whilst visual feedback mechanisms were enhanced to provide immediate response to user adjustments. The layout was optimised to better support extended testing sessions, and progress indicators were refined to help participants maintain awareness of their position within the protocol.

Technical aspects of the platform were also substantially improved. The data validation mechanisms were enhanced to ensure more robust data col-

lection, whilst error handling and recovery procedures were strengthened to maintain system stability. Audio playback controls were optimised to ensure consistent quality across different devices and connection speeds, and database interaction patterns were refined to improve data storage efficiency and reliability.

Procedural elements of the protocol were similarly refined based on participant feedback. Test instructions were clarified to ensure consistent understanding across participants, and environmental requirement checks were standardised to maintain testing conditions. The practice trial implementation was enhanced to better prepare participants for the main tasks, and timing and progression controls were adjusted to optimise the flow between different protocol phases.

Validation Outcomes The comprehensive validation process established that the protocol effectively:

- Captures reliable and meaningful audiovisual associations
- Maintains consistent participant engagement throughout extended testing sessions
- Provides robust data collection across all protocol components
- Supports detailed analysis of emotional components in crossmodal associations
- Demonstrates sensitivity to individual differences in musical and visual abilities

These validation results confirmed the protocol's suitability for large-scale implementation in investigating emotion-mediated audiovisual associations. The refined protocol provides a robust foundation for the subsequent studies presented in this thesis, ensuring methodological rigour and data quality in the investigation of crossmodal perception and emotional processing.

2.5.3 Protocol Structure and Components

The validated protocol comprises four distinct components, each designed to capture specific aspects of audiovisual associations and their emotional correlates. Each component was structured to ensure comprehensive data collection whilst maintaining participant engagement and data quality.

Form The protocol begins with a comprehensive assessment of participants' background and preferences. This initial phase collects:

- Demographic information
- Musical background and training
- Artistic experience and preferences
- Musical genre preferences
- Self-reported emotional responses to specific colours and shapes
- History of any synaesthetic experiences

The full text of the form presented to participants is reported in appendix A.

Pre-test Following the preliminary assessment, participants complete a series of standardised tests to evaluate their visual and musical abilities:

1. **Ishihara Test:** A standardised colour blindness assessment comprising 24 plates, where participants identify numbers or patterns within colour-coded images.
2. **Perfect Pitch Test:** Participants identify musical notes on a two-octave keyboard after a single presentation, with each tone playable only once. The test progresses sequentially through multiple tones.
3. **Melodic Discrimination Test:** Based on [76], this test presents three melodies where:

- Two melodies share identical melodic structure, whereas one has some differences
- The second melody is pitched a semitone higher than the first
- The third melody is pitched a full tone higher
- Participants identify the melodically distinct sequence

4. **Mistuning Perception Test:** Adapted from [77], participants evaluate vocal pitch accuracy across paired musical sequences.
5. **Beat Alignment Test:** Following [78], participants assess rhythmic synchronisation, identifying temporal alignment between musical sequences and superimposed rhythmic elements.

The interface of the pre-test can be found in A.

Phase One: Music-to-Image Generation In this core experimental phase, participants create visual representations while listening to musical stimuli. The image generation interface allows manipulation of multiple parameters:

- Number of visual objects
- Object dimensions
- Spatial dispersion
- Shape morphology (ranging from angular to rounded, following Maluma/-Takete principles [79])
- Colour properties (hue, saturation, brightness)

Musical stimuli were specifically composed for this study by Andrea Sorbo to eliminate potential confounds from musical familiarity. The composition set includes various musical genres, ensuring broad stylistic coverage while maintaining controlled musical parameters. The information

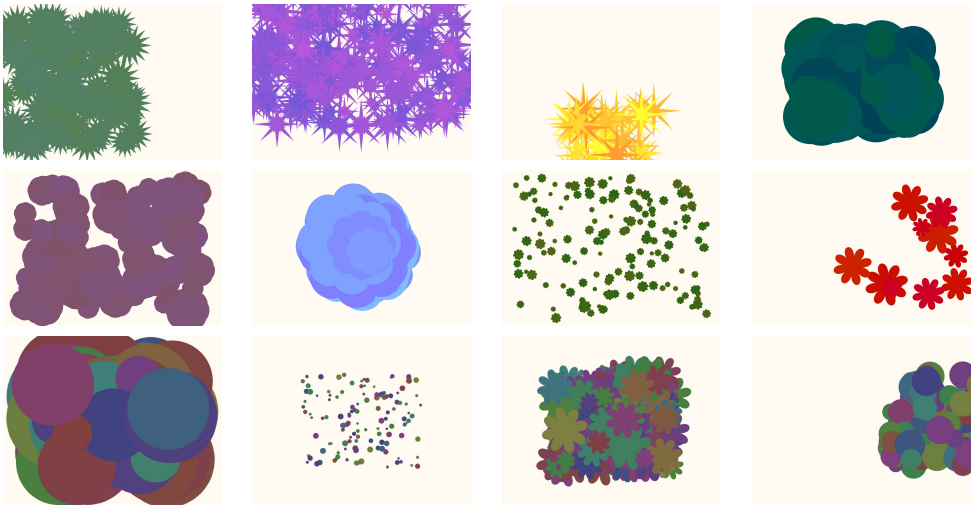


Figure 2.9. Examples of participant-generated images during Phase One, demonstrating variations in visual parameters in response to different musical stimuli. The first two rows show responses to contrasting musical pieces, while the bottom row displays examples of extreme parameter settings.

Table 2.5. Information about musical tracks presented in the first phase of the test.

Track Number	Genre	Key	BPM
1	Choir	Fmin	140
2	POP	Dmaj	140
3	Orchestral Soundtrack	Bmin	94
4	Trap	Cmin	110
5	Electronic	Cmin	128
6	Minimal Soundtrack	Emaj	90
7	Country	Gmaj	74
8	Jazz	Dmaj	90
9	Classical	Bmaj	100
10	Latino	Amin	90

of the generated songs are reported in Table 2.5 and can be played at the following link: https://urly.it/312gn6.

After image generation, participants rate their emotional responses using the 9-term GEMS scale: Amazement, Solemnity, Tenderness, Nostalgia, Calmness, Power, Joyful Activation, Tension, and Sadness.

The protocol allows for:

- Single playback of each musical piece during image creation
- Optional second playback before emotional response recording
- Real-time parameter adjustment with visual feedback
- Submission only after music playback completion

Phase Two: Image-to-Emotion Assessment The final phase presents 21 standardised images where individual parameters are systematically varied while maintaining others at neutral values, as shown in Table 2.6. The parameters that are varied across the 21 images are: 1) Number of visual objects (low = 5, mid = 250, high = 500), 2) Object dimensions (low = 0.5x, mid = 1x, high = 10x, with 0.5x and 10x representing proportional scaling relative to the image size), 3) Spatial dispersion (low = 0.5x, mid = 1x, high = 10x, with 0.5x and 10x representing proportional scaling relative to the image size), 4) Shape morphology (low = spiky, mid = circle, high = rounded flower-like), 5) Hue (low = 20, mid = 70, high = 100), 6) Saturation (low = 20, mid = 70, high = 100), 7) Brightness (low = 20, mid = 70, high = 100). This controlled manipulation allows for isolation of parameter-specific emotional responses. Participants evaluate each image using the same GEMS scale employed in Phase One, enabling direct comparison of emotional responses across modalities. A representation of the visual interface is shown in Figure 2.10.

Each component of the protocol was designed to build upon previous elements, creating a comprehensive assessment of audiovisual associations while maintaining participant engagement and data quality. The structured

Image	Visual Objects	Object Dimensions	Spatial Dispersion	Shape Morphology	Hue	Saturation	Brightness
Neutral Parameters	Mid	Mid	Mid	Mid	Mid	Mid	Mid
Low Visual Objects	Low	Mid	Mid	Mid	Mid	Mid	Mid
High Visual Objects	High	Mid	Mid	Mid	Mid	Mid	Mid
Low Object Dimensions	Mid	Low	Mid	Mid	Mid	Mid	Mid
High Object Dimensions	Mid	High	Mid	Mid	Mid	Mid	Mid
Low Spatial Dispersion	Mid	Mid	Low	Mid	Mid	Mid	Mid
High Spatial Dispersion	Mid	Mid	High	Mid	Mid	Mid	Mid
Spiky Shape	Mid	Mid	Mid	Low	Mid	Mid	Mid
Rounded Shape	Mid	Mid	Mid	High	Mid	Mid	Mid
Low Hue	Mid	Mid	Mid	Mid	Low	Mid	Mid
High Hue	Mid	Mid	Mid	Mid	High	Mid	Mid
Low Saturation	Mid	Mid	Mid	Mid	Mid	Low	Mid
High Saturation	Mid	Mid	Mid	Mid	Mid	High	Mid
Low Brightness	Mid	Mid	Mid	Mid	Mid	Mid	Low
High Brightness	Mid	Mid	Mid	Mid	Mid	Mid	High

Table 2.6. Table of 21 standardized images with systematically varied visual parameters that were used during the second phase of the test in order to collect emotional response to visual stimuli.

progression from basic ability assessment through to complex audiovisual tasks enables detailed analysis of the relationships between musical ability, visual perception, and emotional responses in crossmodal associations.

A schematic representation of the protocol is shown if figure 2.11

2.5.4 Participants

Participant Recruitment and Demographics

The study recruited Italian-speaking volunteers through various channels over a two-month period (February-March 2022). The recruitment strategy involved direct communication to researchers and doctoral candidates, university mailing lists, and flyer distribution on campus. All participants were required to provide written informed consent and agree to the privacy policy in compliance with the European General Data Protection Regulation (GDPR) guidelines. The study protocol was approved by Università Campus Bio-Medico di Roma's ethical committee on February 16, 2022, with number of clinical studies' register 2021.236. The initial demographic data of the enrolled participants (N=123) is summarised in Table

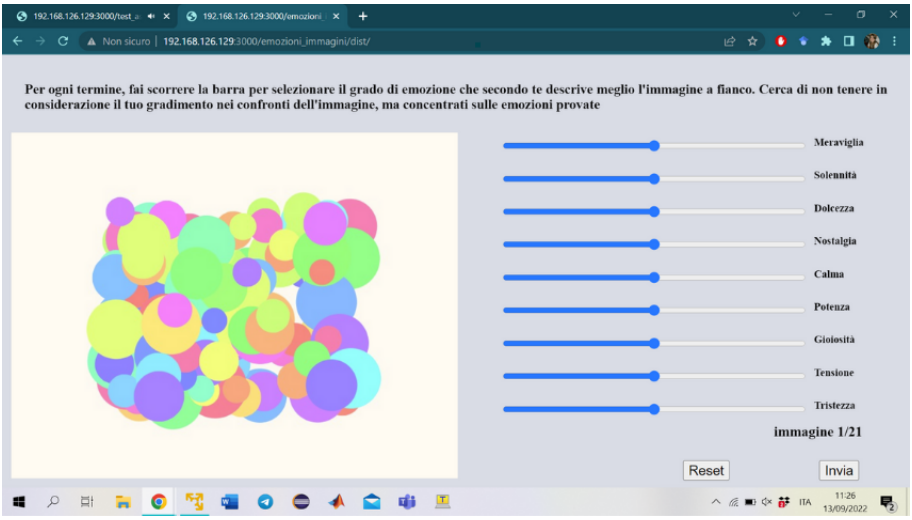


Figure 2.10. Visual interface of the second phase. Participants are shown 21 images and asked to describe their emotional response using the GEMS.

Protocol Summary

Pretest:

Five tests that evaluate visual and musical abilities:

1. Ishihara
2. Perfect Pitch
3. Melodic Discrimination
4. Mistuning Perception
5. Beat Alignment

First Phase:

Objective: Evaluate how music affects the emotions and creative process of individuals.
INPUT: 10 songs belonging to different genres
OUTPUT: 10 Images described by 7 graphical parameters and scores for 9 music-induced emotions

Second Phase:

Objective: Evaluate visually induced emotions
INPUT: 21 Images, each with a graphical parameter set to a high value
OUTPUT: Scores for 9 image-induced emotions

DESCRIPTION OF INPUTS AND OUTPUTS

MUSICAL GENRES:

1. Choir
2. Pop
3. Orchestral Soundtrack
4. Trap
5. Electronic
6. Minimal Soundtrack
7. Country
8. Jazz
9. Classical
10. Latino

GRAPHICAL PARAMETERS:

1. Objects' Number
2. Objects' Dimension
3. Dispersion
4. Shape
5. Hue
6. Saturation
7. Colour

EMOTIONS:

1. Amazement
2. Solemnity
3. Tenderness
4. Nostalgia
5. Calmness
6. Power
7. Joyful Activation
8. Tension
9. Sadness

Figure 2.11. Schematic representation of the test phases. The test was administered remotely, and subjects could complete the pretest and the two main phases in a single or in separate sessions. Subjects belonging to Group A completed all the phases, whereas subjects from Group B skipped phase one.

2.7, which shows the distribution between experimental (n=81) and control (n=42) groups, gender distribution (56 males, 67 females), and age characteristics (range: 19-41 years, mean age=23.4).

The enrolment process occurred in two phases: initially, 84 subjects were recruited and divided into two groups (Group A and Group B) of 42 subjects each. In the second phase, an additional 39 subjects were recruited and added to Group A, bringing its total to 81 subjects. The total enrolled cohort demonstrated a balanced gender distribution (M=56 (45.53%), F=67 (54.47%), SD=4.7). Of the initial cohort, 39 participants did not commence the test after enrolment, and 4 started but did not complete it, resulting in a final analysed sample of 80 participants (M=34 (42.5%), F=46 (57.5%), mean age=23.07, SD=5.3). The age distribution reflected a predominantly young adult population, with ages ranging from 19 to 41 years. During the experiment, 80% of participants completed the test in a single session, while the remaining 20% opted for multiple sessions, with no interval exceeding one week between sessions.

The final composition of the groups was as follows: Group A consisted of 47 subjects (M=22 (46.8%), F=25 (53.2%), mean age=22.34), while Group B comprised 33 subjects (M=15 (45.5%), F=18 (54.5%), mean age=24.12).

Table 2.7. Description of the enrolled subjects that participated to the experiment.

Feature	**Value**
[Total - Male - Female] Number of Subjects	[123 - 56 - 67]
Group [A - B]	[81 - 42]
[Minimum - Average - Maximum] Age	[19 - 23.4 - 41]

Group Assignment Methodology

The methodology for group assignment followed a systematic, data-driven approach comprising several key steps based on the answers given in the form described in the previous section of this chapter:

1. Initial clustering using k-means algorithm to identify natural groupings within the participant population. This step employed a multidimensional approach, considering all the given answers simultaneously to ensure comprehensive grouping.

2. Statistical validation of cluster distinctiveness using Kruskal-Wallis testing ($p < 0.001$). This non-parametric approach was chosen due to its robustness in handling potentially non-normal distributions in psychological data.

3. Randomized allocation of participants from each cluster to form A and B groups.

4. Verification of group equivalence through subsequent Kruskal-Wallis analysis ($p > 0.05$), ensuring that the final groups were statistically comparable across all relevant dimensions.

The k-means clustering algorithm initially identified two significantly different groups, as indicated by the Kruskal-Wallis test results ($p < 0.001$). After reassigning groups based on the steps described above, the Kruskal-Wallis test applied to groups A and B confirmed that the two groups are now uniform.

As described in the group A was designated to complete the full protocol, including all phases, whilst the group B participated only in the pretest and phase two, omitting phase one. Post-enrolment, 47 participants withdrew from the study before test completion, resulting in a final analytical sample of 80 participants (34 males [42.5%] and 46 females [57.5%], mean age 23.07 years, SD=5.3). During the experiment, 80% of participants chose to complete the test in one session, while the remaining participants opted to finish it in multiple sessions, with no interval exceeding one week between sessions. The final composition showed Group A with 47 subjects (22 males [46.8%], 25 females [53.2%], mean age 22.34 years) and Group B with 33 subjects (15 males [45.5%], 18 females [54.5%], mean age 24.12 years).

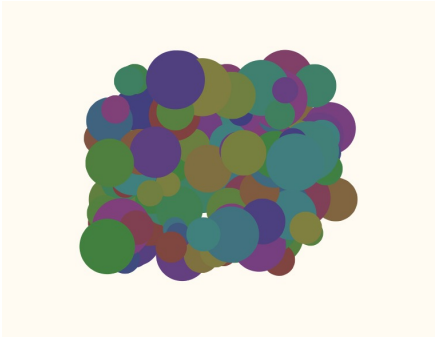
Robustness and impartiality of the results

A comparative analysis of emotional responses elicited by the neutral image across both groups is illustrated in Figure 2.12. Statistical analysis using the Kruskal-Wallis test demonstrated a significant distinction between the groups ($p < 0.01$), suggesting that Phase 1 participation influenced the participants' emotional state. The impact was particularly evident in emotions such as Tenderness, Calmness, Power and Joyful Activation, as depicted in the boxplot in Figure 2.12.

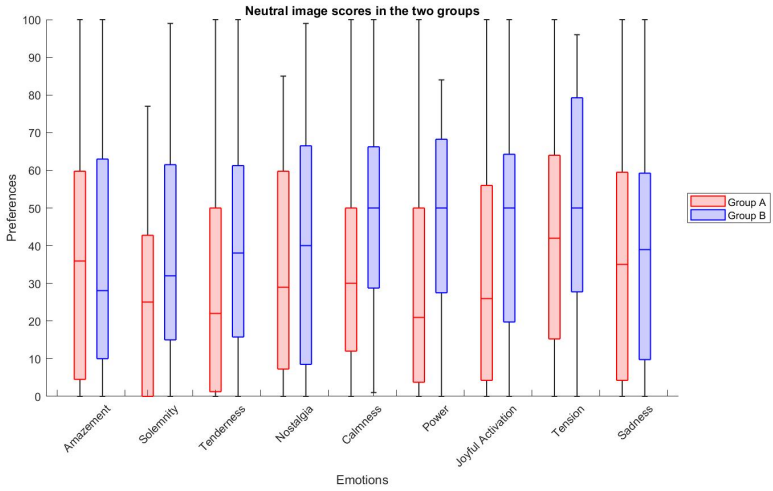
To account for baseline differences, each participant's emotional ratings for the neutral image were subtracted from their responses to other images. When comparing these baseline-corrected emotional variations between groups, no significant differences were observed ($p > 0.05$). This indicates an absence of familiarity bias in the test, whilst confirming that the altered emotional baseline does not influence specific associations once baseline differences are controlled for. Figure 2.13 presents comprehensive statistical results for each image.

2.5.5 Quantifying Emotional Mediation in Audiovisual Crossmodal Parameters

To quantify the role of emotions in mediating audiovisual crossmodal associations, we conducted a two-stage analysis focusing exclusively on Group B participants to avoid any potential familiarity bias. The first stage examined the uniformity of association patterns across different population subgroups, investigating whether distinct clusters of individuals exhibit systematically different approaches to emotional-visual associations. The second stage assessed Emotional correspondence across the entire Group B dataset, quantifying how music-induced emotions influence visual perception and parameter selection. These complementary analyses, detailed in the following sections, provide a comprehensive understanding of the emotional mechanisms underlying audiovisual crossmodal associations.



(a)



(b)

Figure 2.12. The neutral image presented to Groups A and B as an emotional state reference at the onset of phase 2 is shown in (a). This image was constructed with median values for parameters including saturation, brightness, dimension, dispersion, shape and numerosity, whilst hue varied across objects, encompassing the complete 360° colour wheel spectrum. The boxplot in (b) contrasts the emotional responses between Group A participants, who had completed phase one prior to viewing the neutral image, and Group B participants, who had not undertaken phase one. The assessment scale spans from 0 to 100, where 100 signifies strong preference and 0 indicates minimal preference. The data suggests that phase one completion altered Group A's emotional baseline.

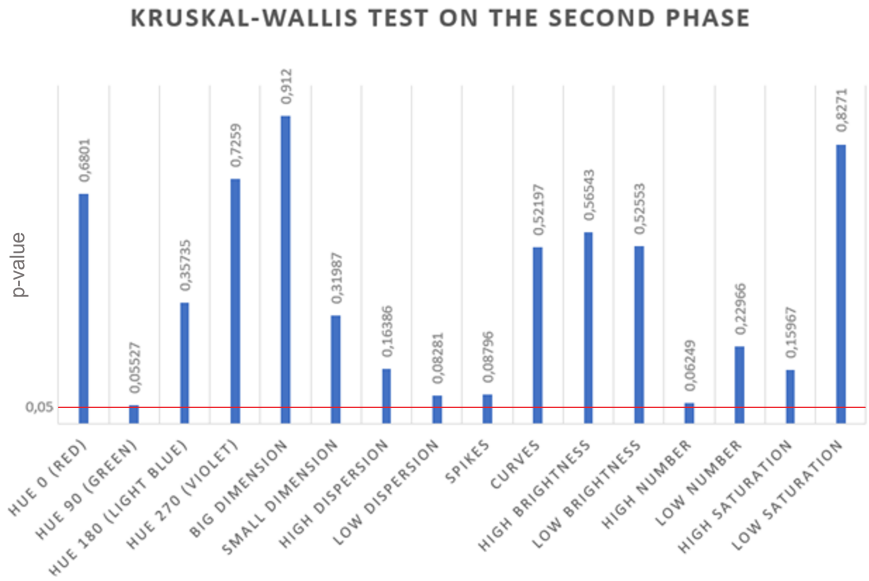


Figure 2.13. Differential emotional responses between Groups A and B during phase 2 were evaluated using the Kruskal-Wallis test. During this phase, participants viewed images with median values across all parameters, save for one parameter set to an extreme value. Emotional responses were calculated as deviations from each participant's neutral image response. The bars represent p-values, and the analysis indicated no statistically significant differences between the groups ($p>0.05$).

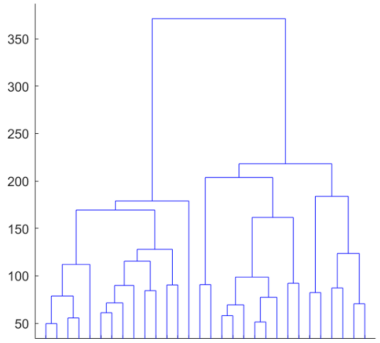
Population Subgroups Analysis

The first stage of our analysis investigated whether different population subgroups employ distinct strategies in audiovisual associations. This investigation aimed to verify if the association process remains consistent across all subjects or if different population groups require distinct modeling approaches. We analysed data from Phase Two to identify potential clusters with different emotion-to-image association logics, then examined their Phase One responses for systematic differences. The analysis followed a systematic six-step process:

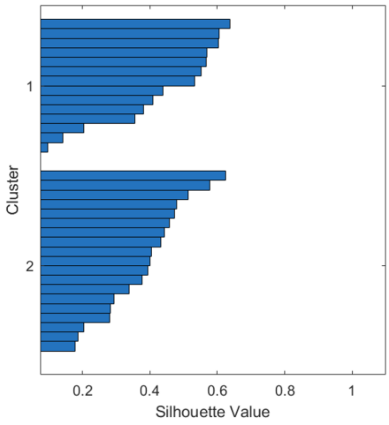
Step 1. Division of Subjects into Clusters Initial clustering was performed using Agglomerative hierarchical cluster tree analysis on Phase Two data. For each image, we generated two, three, and four cluster configurations to identify optimal groupings.

Step 2. Cluster Compactness Computation For each clustering configuration:

- Calculated silhouette scores for individual clusters
- Computed average silhouette values to measure overall cluster response compactness
- Selected configurations with highest compactness for each image



(a) Example of agglomerative hierarchical cluster tree results.



(b) Example of silhouette plot.

Figure 2.14. Clustering analysis visualisations.

Step 3. Selection of Optimal Clustering For each identified clustering:

- Calculated compactness for image-emotion associations
- Computed average compactness indices

- Applied ANOVA tests to Phase Two results
- Averaged test values for group diversity index
- Selected final clustering based on:
  - High compactness
  - Low ANOVA p-values
  - Balanced subject distribution between groups

Table 2.8. Results of clustering analysis across different parameters. The cluster based on the emotional reactions to the neutral image was the selected one, as it has the highest mean compactness and lowest mean p ANOVA.

Parameter	Mean Compactness	Mean p ANOVA	Cluster Distribution
Neutral	0.31	1.282e-08	C1: 21 - C2: 26
Green	-0.11	0.043	C1: 5 - C2: 15 - C3: 27
Red	0.11	9.549e-07	C1: 20 - C2: 27
Light Blue	0.090	2.795e-05	C1: 15 - C2: 32
Violet	0.010	9.232e-08	C1: 16 - C2: 11 - C3: 20
Big Dimension	0.12	0.000031	C1: 18 - C2: 29
Small Dimension	0.040	1.526e-07	C1: 18 - C2: 29
High Dispersion	0.18	1.097e-06	C1: 18 - C2: 29
Low Dispersion	0.060	6.988e-08	C1: 23 - C2: 24
Spikes	0.16	7.290e-08	C1: 15 - C2: 32
Curves	0.26	3.669e-05	C1: 12 - C2: 35
High Brightness	0.22	4.497e-05	C1: 7 - C2: 40
Low Brightness	0.19	4.874e-08	C1: 14 - C2: 33
High Number	0.12	3.124e-08	C1: 4 - C2: 21 - C3: 22
Low Number	0.010	7.371e-08	C1: 11 - C2: 23 - C3: 13
High Saturation	0.29	1.531e-08	C1: 8 - C2: 39
Low Saturation	0.15	1.382e-05	C1: 11 - C2: 36

Step 4. Computation of Emotional correspondence For each cluster, we computed Emotional correspondence using:

$$\text{emotional consistency} = 1 - \frac{|A(i) - B(i)|}{100} \tag{2.1}$$

where i represents the emotion, A represents emotional spectra from Phase One, and B represents emotional spectra from Phase Two.

	Green	Red	Blue	Purple	Dim+	Dim-	Disp+	Disp-
C1	0.670	3.52	2.14	1.76	5.24	7.24	1.48	8.57
C2	0.920	3.42	2.27	1.69	5.77	7.96	1.12	8.92

	Spikes	Curves	Lum+	Lum-	Num+	Num-	Sat+	Sat-
C1	2.48	5.00	6.90	4.71	4.86	4.00	1.90	8.57
C2	1.85	5.81	7.04	5.58	5.35	3.96	1.92	10.0

Table 2.9. Number of generated images with high parameter values divided by group size.

Step 5. Group Coherence Analysis We applied the Kruskal-Wallis test to compare coherence between groups. Results with $p < 0.01$ indicated significant differences in coherence patterns.

Step 6. Musical Feature Analysis We extracted twelve features from each musical piece using the MIRAudio toolbox in MATLAB, including:

- Mirzerocross: Signal noise measurement
- Mircentroid: Spectral distribution centre
- Mirspread: Spectral distribution width
- Additional features detailed in Table 2.12

Table 2.10. Emotional correspondence values for Group 1.

Parameter	Amazement	Solemnity	Tenderness	Nostalgia	Calmness	Power	Joyful Act.	Tension	Sadness
Green	0.62	0.59	0.77	0.71	0.65	0.67	0.67	0.61	0.62
Red	0.71	0.67	0.64	0.68	0.71	0.56	0.71	0.68	0.71
Light Blue	0.63	0.72	0.80	0.80	0.69	0.75	0.45	0.60	0.63
Purple	0.73	0.80	0.69	0.72	0.61	0.72	0.54	0.76	0.73
Big Dimension	0.73	0.69	0.77	0.64	0.72	0.72	0.59	0.74	0.73
Small Dimension	0.65	0.69	0.72	0.71	0.69	0.58	0.65	0.72	0.65
High Dispersion	0.66	0.81	0.59	0.77	0.59	0.78	0.74	0.73	0.66
Low Dispersion	0.71	0.75	0.72	0.67	0.73	0.65	0.67	0.73	0.71
Spikes	0.69	0.72	0.59	0.66	0.72	0.71	0.74	0.68	0.69
Curves	0.73	0.71	0.69	0.73	0.68	0.72	0.71	0.64	0.73
High Brightness	0.78	0.73	0.68	0.73	0.67	0.73	0.59	0.75	0.78
Low Brightness	0.59	0.59	0.61	0.71	0.68	0.67	0.62	0.74	0.59
High Number	0.69	0.61	0.61	0.72	0.59	0.71	0.51	0.55	0.69
Low Number	0.74	0.76	0.74	0.69	0.75	0.73	0.65	0.71	0.74
High Saturation	0.76	0.79	0.78	0.68	0.76	0.78	0.60	0.85	0.76
Low Saturation	0.72	0.74	0.66	0.76	0.67	0.74	0.64	0.59	0.72

Table 2.11. Emotional correspondence values for Group 2.

Parameter	Amazement	Solemnity	Tenderness	Nostalgia	Calmness	Power	Joyful Act.	Tension	Sadness
Green	0.74	0.59	0.77	0.63	0.60	0.62	0.66	0.57	0.51
Red	0.63	0.68	0.67	0.72	0.69	0.61	0.62	0.62	0.67
Light Blue	0.64	0.69	0.72	0.75	0.68	0.66	0.60	0.66	0.75
Purple	0.67	0.59	0.71	0.64	0.57	0.73	0.60	0.59	0.65
Big Dimension	0.71	0.65	0.78	0.59	0.65	0.69	0.59	0.66	0.65
Small Dimension	0.63	0.68	0.59	0.72	0.61	0.61	0.59	0.58	0.66
High Dispersion	0.66	0.77	0.79	0.67	0.59	0.59	0.58	0.81	0.74
Low Dispersion	0.59	0.73	0.59	0.66	0.69	0.67	0.68	0.72	0.75
Spikes	0.66	0.62	0.72	0.67	0.64	0.65	0.74	0.60	0.76
Curves	0.71	0.72	0.73	0.71	0.57	0.64	0.64	0.58	0.68
High Brightness	0.75	0.72	0.69	0.71	0.66	0.68	0.67	0.71	0.76
Low Brightness	0.69	0.67	0.66	0.65	0.66	0.68	0.56	0.75	0.68
High Number	0.66	0.65	0.61	0.59	0.53	0.68	0.50	0.52	0.56
Low Number	0.59	0.69	0.77	0.66	0.67	0.67	0.63	0.67	0.74
High Saturation	0.73	0.81	0.74	0.66	0.75	0.74	0.58	0.85	0.80
Low Saturation	0.72	0.72	0.71	0.68	0.60	0.67	0.62	0.68	0.69

The analysis revealed minimal differences between groups. Statistical significance was found in only 8 out of 144 emotion-graphical parameter pairs (5.6%), which can be attributed to multiple comparison effects rather than genuine group differences. Similarly, PCA analysis of musical features showed no distinct clustering patterns, with only minor variations

Table 2.12. Musical features extracted for analysis.

Feature	Description
Mirzerocross*	Counts the number of times the signal crossed the X-axis, indicating noisiness
Mircentroid	Returns the centroid of the data. It is used to describe the spectral distribution
Mirspread	Returns the standard deviation of the data. It is used to describe the spectral distribution
Mirentropy	Returns the relative Shannon entropy of the input, computed as H(p)=-sum(p.*log(p))/log(length(p)). It offers a general description of the input curve p, and indicates in particular whether it contains predominant peaks or not
Mirflatness	Returns the flatness of the data, which indicates whether the distribution is smooth or spiky, and results from the simple ratio between the geometric mean and the arithmetic mean
Mirroloff*	Finds the frequency such that a certain fraction of the total energy is contained below that frequency. This ratio is fixed by default to 0.85 to estimate the amount of high frequency in the signal
Mirlowenergy*	Computes the low energy rate, i.e. the percentage of frames showing less-than-average energy. This measure is used to estimate the temporal distribution of energy
Mirmode**	Estimates the modality, i.e. major vs. minor, returned as a numerical value between -1 and +1
Mirkey (1)**	Gives a broad estimation of tonal centre positions by returning the best key(s), i.e., the peak abscissa(e)
Mirkey(2)**	Gives a broad estimation of the tonal centre positions' respective clarity by computing the key clarity
Mirrms*	Computes the global energy of the signal x by taking the root average of the square of the amplitude
Mirroughness	Estimates the total roughness by computing the peaks of the spectrum and taking the average of all the dissonance between all possible pairs of peaks

* Performs frame decomposition with 50 ms frame length and half overlapping
** Performs frame decomposition with 1 s frame length and 50% hop factor (0.5 s)

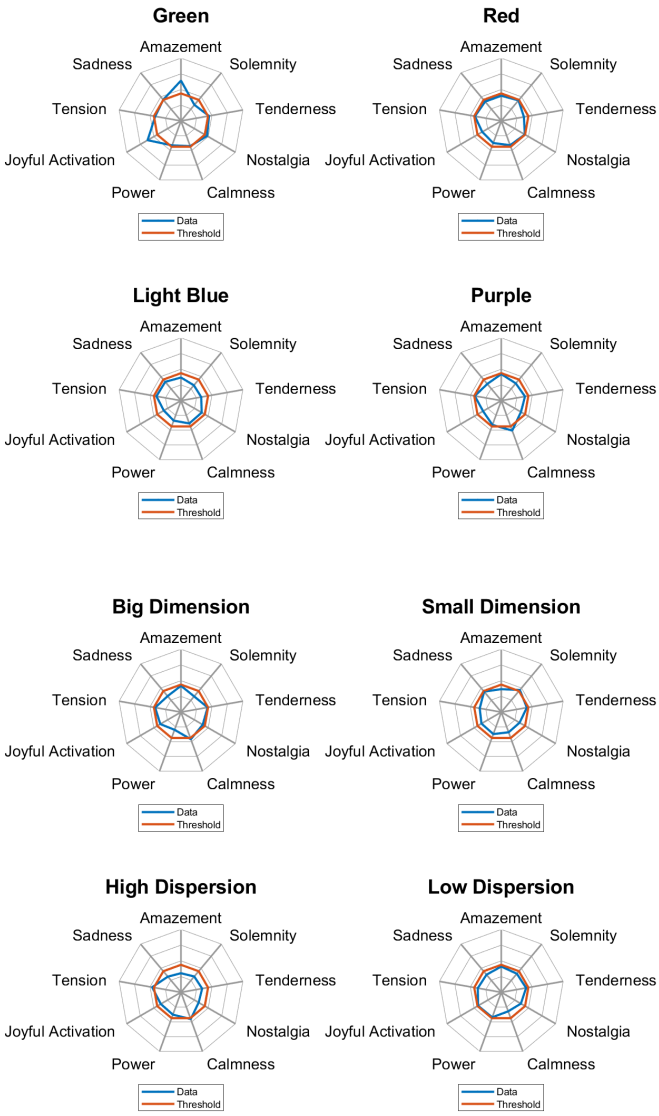


Figure 2.15. ECDF analysis results (Part 1) showing consistency distribution across emotional dimensions. Values closer to centre indicate higher consistency.

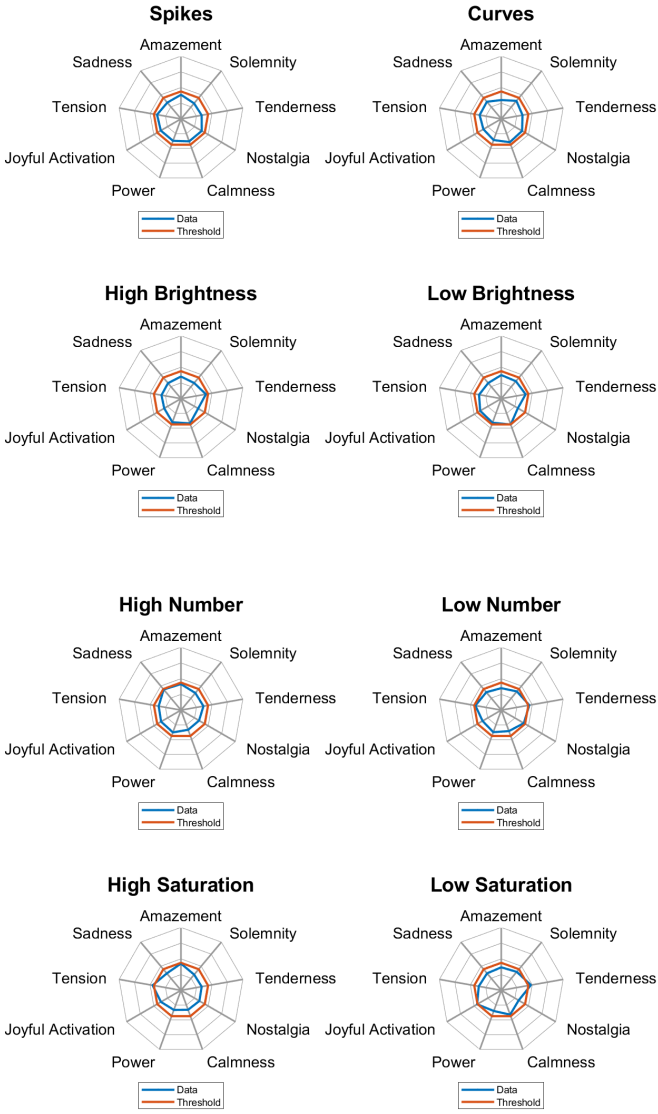
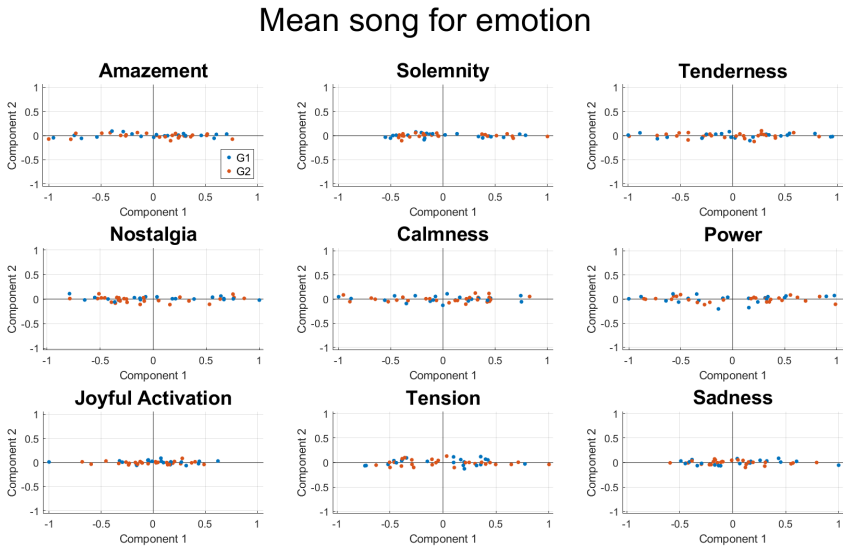
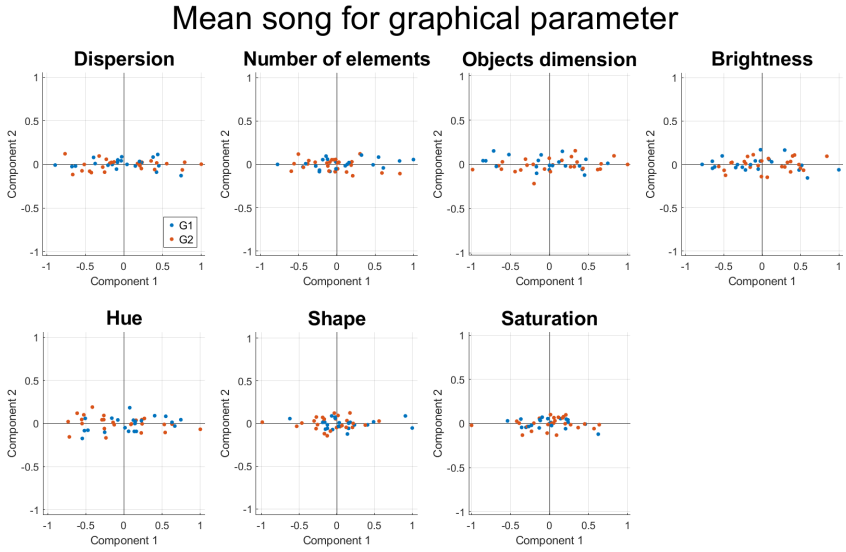


Figure 2.16. ECDF analysis results (Part 2) showing consistency distribution across emotional dimensions. Values closer to centre indicate higher consistency.

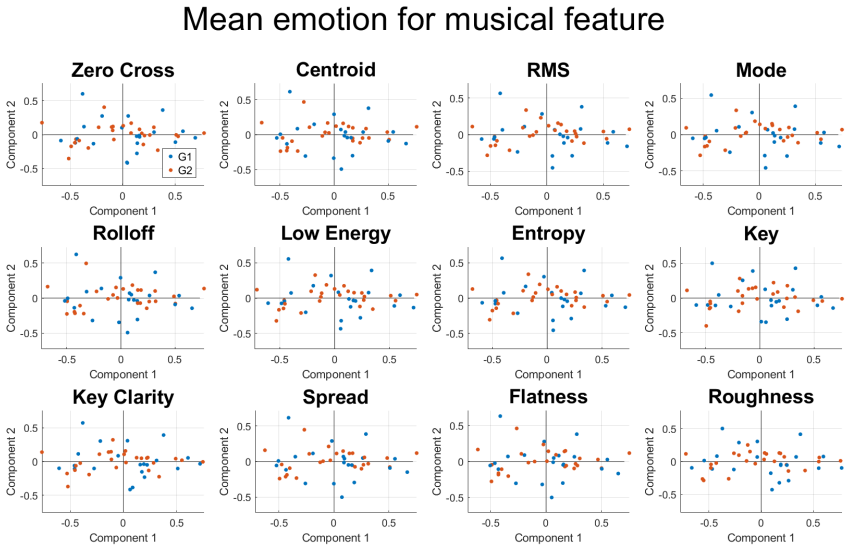


(a) First PCA analysis of musical features

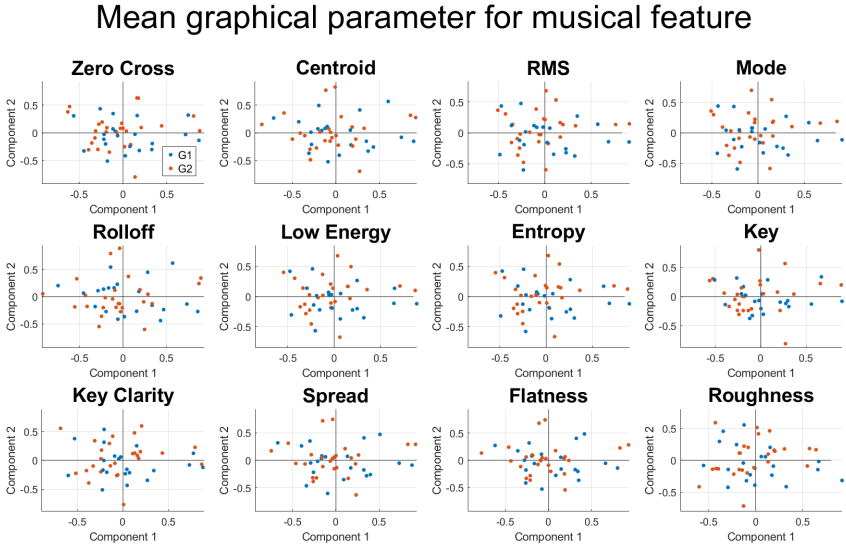


(b) Second PCA analysis of musical features

Figure 2.17. Principal Component Analysis results (Part 1) showing no distinct clustering patterns between groups.



(a) Third PCA analysis of musical features



(b) Fourth PCA analysis of musical features

Figure 2.18. Principal Component Analysis results (Part 2) showing no distinct clustering patterns between groups.

in Beat Alignment Test scores ($p = 0.0133$) and pop music preferences ($p = 0.0298$).

Quantifying Emotional Mediation Through Consistency Analysis

Following the analysis of population subgroups, we investigated the role of emotions in mediating audiovisual associations across the entire Group B dataset. This analysis aimed to quantify how emotions induced by music during image generation correlate with emotions evoked by viewing similar images. The analysis assessed the Emotional correspondence between Phase One and Phase Two responses by:

1. Computing mean values for each graphical parameter per subject
2. For each emotion-parameter pair:
    - Identifying Phase One songs where the generated image's parameter exceeded the mean
    - Calculating Emotional correspondence using Equation 2.1 presented earlier

A low Emotional correspondence value indicates disagreement between music-induced emotions in Phase One and image-induced emotions in Phase Two, suggesting minimal emotional mediation. Conversely, high consistency suggests strong emotional mediation in the audiovisual association 2.13. Three scenarios can lead to high consistency:

1. High values in both phases: the music evokes strong emotion leading to high parameter values, and viewing images with high parameter values later evokes the same strong emotion
2. Low values in both phases: minimal emotional response to music corresponds to low parameter values, with similar low emotional response to viewing such images
3. Medium values in both phases: moderate emotional responses consistently appear in both music and image perception

Table 2.13. Examples of Emotional correspondence calculations for different phase combinations.

Value of emotion in Phase 1	Value of graphical parameter in Phase 1	Value of emotion in Phase 2	Emotional correspondence index
100	70	100	1
0	10	0	1
50	100	50	1
0	20	80	0.2
20	80	70	0.5

The analysis revealed several significant patterns of emotional mediation in audiovisual associations. Table 2.14 presents the comprehensive results of Emotional correspondence calculations across all parameters and emotions.

Parameter	Amazement	Solemnity	Tenderness	Nostalgia	Calmness	Power	Joyful Act.	Tension	Sadness
Green	0.59	0.59	0.77	0.71	0.65	0.67	0.67	0.61	0.62
Red	0.71	0.67	0.64	0.68	0.71	0.56	0.71	0.68	0.71
Light Blue	0.63	0.72	0.80	0.80	0.69	0.75	0.45	0.60	0.63
Purple	0.73	0.80	0.69	0.72	0.61	0.72	0.54	0.76	0.73
Big Dimension	0.73	0.69	0.77	0.64	0.72	0.72	0.70	0.74	0.73
Small Dimension	0.65	0.69	0.72	0.71	0.69	0.58	0.65	0.72	0.65
High Dispersion	0.66	0.81	0.70	0.77	0.70	0.78	0.74	0.73	0.66
Low Dispersion	0.71	0.75	0.72	0.67	0.73	0.65	0.67	0.73	0.71
Spikes	0.69	0.72	0.70	0.66	0.72	0.71	0.74	0.68	0.69
Curves	0.73	0.71	0.69	0.73	0.68	0.72	0.71	0.64	0.73
High Brightness	0.78	0.73	0.68	0.73	0.67	0.73	0.70	0.75	0.78
Low Brightness	0.70	0.70	0.61	0.71	0.68	0.67	0.62	0.74	0.70
High Number	0.69	0.61	0.61	0.72	0.59	0.71	0.51	0.55	0.69
Low Number	0.74	0.76	0.74	0.69	0.75	0.73	0.65	0.71	0.74
High Saturation	0.76	0.79	0.78	0.68	0.76	0.78	0.60	0.85	0.76
Low Saturation	0.72	0.74	0.66	0.76	0.67	0.74	0.64	0.70	0.72

Table 2.14. Emotional correspondence indices between visual parameters and emotions across experimental phases. Higher values indicate stronger emotional mediation in audiovisual associations.

Key findings from the consistency analysis include:

- Sadness showed the highest consistency indices, particularly with:
  - Low Dispersion (0.78)
  - High and Low Brightness (0.76 for both)
  - Low Saturation (0.75)
- Amazement demonstrated strong consistency with:
  - Curvilinear shapes (0.75)
  - Both curvilinear and spiked shapes
  - Low dispersion and saturation parameters
- Tenderness showed notable consistency patterns:
  - Strong association with green hues (0.75)
  - Numerosity emerged as the most dichotomic parameter
- Additional significant associations included:
  - Nostalgia influencing both shape types (0.70-0.73) and dimensional parameters (0.71-0.72)
  - Power and Solemnity showing consistent relationships with saturation and brightness
  - Tension demonstrating strong influence on brightness perception (0.72-0.74)

These findings align with previous research on emotion-mediated crossmodal associations, while providing novel quantitative insights into the specific roles of different emotions in mediating audiovisual parameters. The high consistency values across multiple parameter-emotion pairs suggest that emotional responses play a crucial role in bridging auditory and visual perception, particularly in areas such as colour perception, shape recognition, and spatial organisation. The analysis demonstrates that emotional mediation in audiovisual associations is not only present but quantifiable

and relatively consistent across subjects. This understanding provides valuable insights for applications in fields ranging from multimedia design to sensory substitution technologies, where emotional coherence between auditory and visual elements is crucial for effective communication and experience design.

Insights on Music-Induced Emotions

To validate our experimental design, we first verified that the selected musical pieces effectively elicited the full range of target emotions. Figure 2.19 shows the mean and standard deviation of emotional scores across all songs, demonstrating comprehensive coverage of the emotional spectrum for all nine GEMS dimensions. To further understand the relationship between musical features and induced emotions, we computed Kendall's correlation coefficients between musical parameters and emotional responses. Musical features were extracted using the MIRaudio toolbox [80], including Zero-cross, Centroid, Spread, Entropy, Flatness, Rolloff, Low Energy, Mode, Key and Key Clarity, RMS, and Roughness.

Emotions	mirmode	flatness	spread	mirroffoff	mircentroid	mirentropy	mirkey	mirrms	mirzerocross	mirlowenergy	key clarity	roughness
Solemnity	-0.27	-0.31	-0.33	-0.29	-0.25	-0.25	-0.14	-0.14	-0.06	-0.01	0.29	0.32
Tension	-0.22	-0.18	-0.20	-0.17	-0.14	-0.18	-0.12	-0.08	-0.04	-0.04	0.20	0.21
Power	-0.18	-0.18	-0.19	-0.17	-0.14	-0.16	-0.15	-0.06	-0.06	-0.11	0.15	0.23
Amazement	-0.06	-0.10	-0.10	-0.09	-0.07	-0.06	-0.08	-0.04	-0.03	-0.04	0.08	0.12
Sadness	-0.17	-0.02	-0.07	-0.06	-0.06	-0.07	-0.02	-0.05	0.04	0.09	0.19	-0.03
Nostalgia	-0.13	0.06	-0.00	0.01	0.00	0.04	-0.02	-0.05	0.09	0.08	0.18	-0.14
Tenderness	0.11	0.01	-0.01	-0.02	-0.05	-0.00	0.13	-0.01	-0.03	0.16	-0.10	-0.12
Calmness	0.12	0.07	0.08	0.06	0.05	0.11	0.05	0.01	0.03	0.03	-0.06	-0.15
Joyful Activation	0.25	0.17	0.22	0.19	0.18	0.19	0.13	0.05	0.02	-0.05	-0.26	-0.12

Table 2.15. Correlation coefficients between musical features and emotions induced during Phase One. Notable correlations include Mode's negative relationship with Solemnity and positive relationship with Joyful Activation.

Key correlations included:

- Mode showed negative correlations with Solemnity and Tension, while positively correlating with Joyful Activation

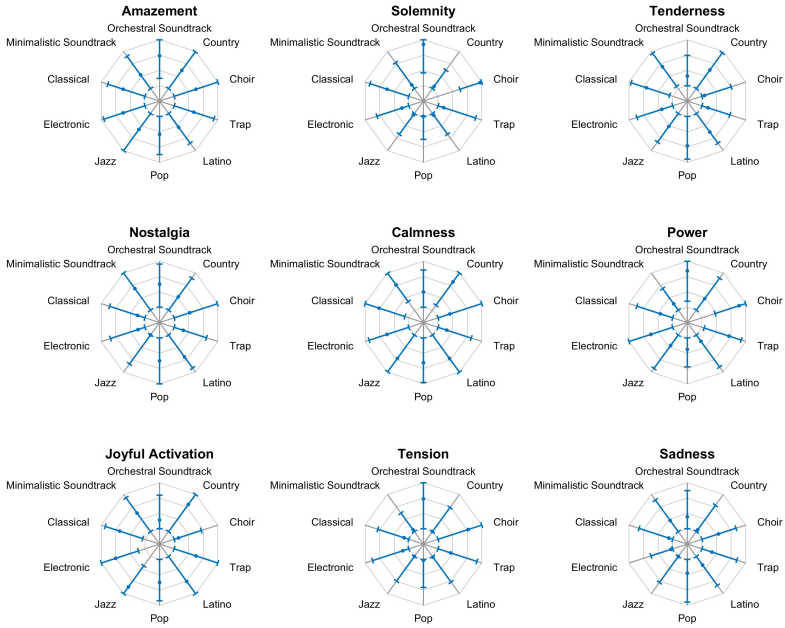


Figure 2.19. Distribution of emotional responses across musical stimuli. Mean and standard deviation of scores for each emotion across the ten musical pieces, demonstrating comprehensive coverage of the emotional spectrum.

- Spectral features (Flatness, Rolloff, Centroid, Entropy) negatively correlated with Solemnity
- Key clarity positively correlated with Solemnity and Tension, while negatively correlating with Joyful Activation
- Roughness showed positive correlations with Solemnity, Tension, and Power

We also examined how musical abilities influence emotional perception by analysing correlations between pretest scores and emotional responses (Figure 2.16).

Emotions	MPT	BAT	MDT
Solemnity	-0.21	-0.12	0.17
Tension	-0.10	-0.04	0.18
Power	-0.07	0.03	0.14
Amazement	-0.07	-0.03	0.04
Sadness	-0.05	-0.17	0.16
Nostalgia	-0.05	-0.22	0.12
Tenderness	-0.02	-0.11	-0.15
Calmness	0.01	-0.04	-0.14
Joyful Activation	0.09	0.20	-0.20

Table 2.16. Correlation coefficients between pretest scores and music-induced emotions. MPT = Mistuning Perception Test, BAT = Beat Alignment Test, MDT = Melodic Discrimination Test.

Notable relationships included negative correlations between Solemnity and mistuning perception, and between Joyful Activation and melodic discrimination. Beat alignment ability showed positive correlation with Joyful Activation and negative correlation with Nostalgia.

Summary of Emotional Mediation Analysis

The comprehensive analysis of emotion-mediated audiovisual associations revealed two key findings. First, through rigorous clustering analysis, we demonstrated that the process of audiovisual association remains remarkably consistent across the population, with only minor variations attributable to individual differences in musical ability and preferences. This uniformity suggests that a single modeling approach can effectively capture these associations. Second, the Emotional correspondence analysis provided quantitative evidence for the mediating role of emotions in audiovisual associations. The high consistency values observed for specific emotion-parameter pairs, particularly for emotions like Sadness and

Amazement, indicate that emotional responses serve as robust bridges between auditory and visual perception. These findings not only validate previous research on crossmodal associations but also provide a quantitative framework for understanding how emotions influence specific visual parameters during audiovisual processing. These results lay the groundwork for the development of predictive models in audiovisual associations, which will be explored in detail in the next section. They also suggest practical applications in fields ranging from multimedia design to assistive technologies, where understanding and leveraging emotional mediation could enhance the effectiveness of audiovisual communication.

2.5.6 Bidirectional Predictive Modeling of Emotion-Based Audiovisual Associations

This investigation examines the bidirectional relationship between emotional responses elicited by musical stimuli and the generation of visual imagery. Our analysis addresses two fundamental research questions that explore the complex interactions between auditory input, emotional experience, and visual perception. The diagram in Figure 2.20 depicts the core

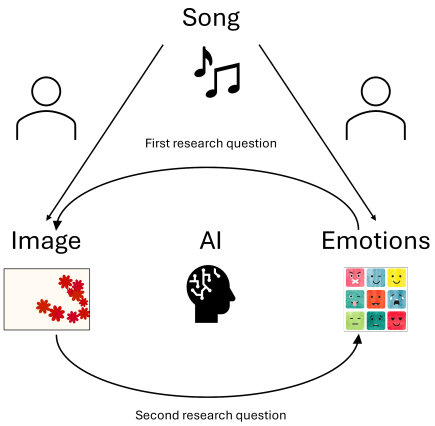


Figure 2.20. Schematic representation of research questions and their interconnections

elements of our investigation and their interrelationships. It illustrates how our research questions examine the connections between musical pieces, visual representations, and emotional responses, with artificial intelligence serving as the analytical framework connecting these components.

Research Question 1: From Emotions to Visual Imagery The first research question evaluates the potential for emotional responses to musical stimuli to serve as predictors for visual imagery characteristics. We examine how accurately visual parameters can be predicted based on participants' emotional responses to musical input. This analysis employs Support Vector Machines (SVM) and Support Vector Regression (SVR) methodologies. Through this approach, we assess both the existence of predictable patterns and the variability in predictability across different visual parameters. This investigation aims to illuminate the mechanisms through which emotional experiences influence visual perception and creation.

Research Question 2: From Visual Imagery to Emotions The second research question investigates the inverse relationship, examining whether computer-generated images can be utilized to determine the emotional responses that influenced their creation. This analysis implements a transformer-based model designed to extract emotional content from visual input. The investigation focuses on the model's ability to predict the emotional responses. This research direction seeks to understand the emotional information embedded in visual creations and evaluate the potential for automated systems to decode the emotional significance of imagery.

Analysis Methods

To examine the relationships between emotional responses and visual parameters, we employed several analytical approaches utilising Python. The investigation centred on two key objectives:

1. Determining whether emotional responses could predict graphical parameter values

2. Assessing whether generated images could predict emotional responses

Forecasting Image Values from Emotional Responses

Our first objective employed three distinct analytical frameworks:

1. **Regression Analysis** - We implemented Support Vector Regression (SVR) to generate continuous value predictions for each graphical parameter based on emotional inputs. Model assessment utilised Mean Absolute Error (MAE).

2. **Three-Category Classification** - We segmented image values into three distinct bands (low, medium, high) and employed Support Vector Machines (SVM) for classification, with accuracy serving as the performance metric.

3. **Binary Classification** - We categorised image values into two groups (low, high), again employing SVM methodology with accuracy measurements for performance assessment.

Each framework was implemented across three contexts:

- A unified model incorporating all musical pieces
- Discrete models for individual musical pieces
- Models based on musical piece clusters

This methodology yielded nine distinct modelling approaches. Each approach generated separate predictions for individual graphical parameters, with performance metrics detailed in the Results subsection.

Musical Piece Clustering Our clustering analysis sought to identify patterns within the musical pieces. Initially, we extracted features via MIRtoolbox in MATLAB, examining the parameter previously reported in table 2.12.

We then applied hierarchical clustering utilising Ward's method, which yielded two distinct musical groupings. The characteristics of these clusters are presented in Table 2.17.

Table 2.17. Characteristics of Musical Track Clusters

Feature	Mean Cluster 1	Std Cluster 1	Mean Cluster 2	Std Cluster 2
Zero Crossing Rate	-0.50954	0.71916	0.3397	1.0693
Spectral Centroid	-0.97436	0.47972	0.64957	0.62935
RMS Energy	-0.50596	0.41481	0.3373	1.1642
Mode	-0.24061	0.91267	0.1604	1.1059
Spectral Rolloff	-1.0171	0.30841	0.67804	0.60309
Low Energy	-0.082068	1.374	0.054712	0.81134
Entropy	-0.78958	0.752	0.52639	0.79338
Key	-0.048983	1.2327	0.032655	0.94078
Key Strength	0.19365	1.118	-0.1291	1
Spectral Spread	-0.081703	1.6965	0.054468	0.25355
Spectral Flatness	-1.0658	0.54415	0.71054	0.32822
Roughness	-0.90089	0.16406	0.60059	0.83769

Deriving Emotional Responses from Generated Images

For our second objective, we employed a transformer-based methodology. This approach utilised a pre-trained transformer model for image analysis, subsequently fine-tuned for our specific task of emotional response prediction. Based on the image analysis, we implemented regression to predict values for each emotional response. We employed Mean Absolute Error (MAE) as the performance metric for this model.

Algorithm Selection Process

Through comparative testing of various algorithms, including Random Forest Regressor and Linear Regressor for the first tasks, and Random Forest and XGBoost for the second task:

- For regression tasks: SVR demonstrated optimal performance

- For classification tasks: SVM yielded superior results

The second objective utilised a transformer-based approach for both image analysis and emotional response prediction. Comprehensive results and performance metrics for all models are presented in the subsequent Results subsection.

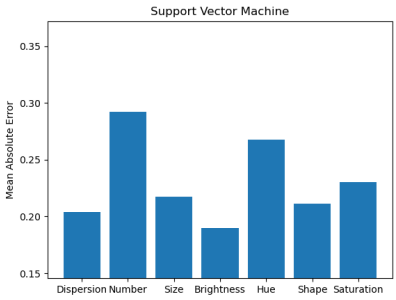
Results

Our analysis yielded findings addressing both primary objectives: forecasting image parameters from emotional responses and deriving emotional responses from generated images.

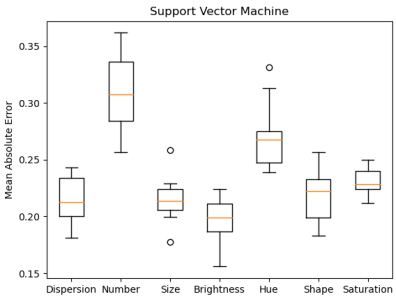
Forecasting Image Parameters from Emotional Responses We evaluated nine distinct modelling approaches, each assessed for their capability to forecast image parameters based on emotional responses to musical stimuli.

Regression Analysis Results SVR methodology enabled prediction of continuous values for each graphical parameter. The regression analyses revealed varying levels of predictive capability across different image parameters. Brightness demonstrated the lowest Mean Absolute Error (MAE) at approximately 0.19, indicating superior predictability amongst all parameters. Dispersion, Shape, and Saturation followed with MAEs ranging from 0.21 to 0.23. Number and Hue proved most challenging to forecast, with MAEs of approximately 0.29 and 0.27 respectively.

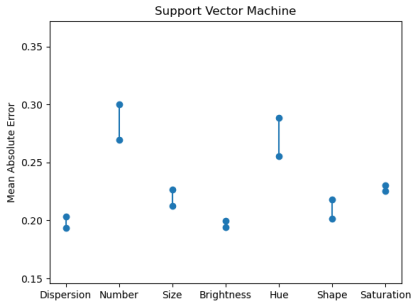
Three-Category Classification Results We employed SVMs to categorise image values into three bands (low, medium, high). The three-category classification demonstrated marginally reduced overall accuracy compared with binary classification. Dispersion, Brightness, and Saturation maintained highest prediction accuracy, ranging from 60% to 70%. Other parameters demonstrated lower accuracy levels, spanning 40% to 50%.



(a) Comprehensive Dataset



(b) Individual Pieces



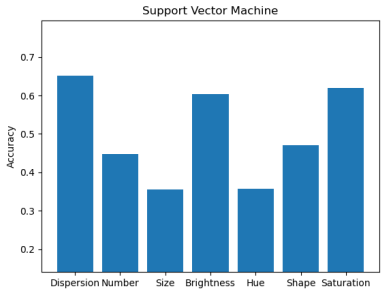
(c) Clustered Pieces

Figure 2.21. SVR model performance in forecasting image parameters across different analytical approaches.

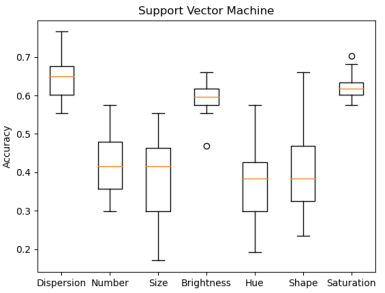
Binary Classification Results SVM methodology was applied for binary classification of image values (low, high).

In binary classification, Dispersion and Saturation achieved highest prediction accuracy, at approximately 85% and 80% respectively. Brightness also demonstrated strong predictability, achieving approximately 80% accuracy. Number, Size, Hue, and Shape proved more challenging to predict, with accuracy ranging from 55% to 60%.

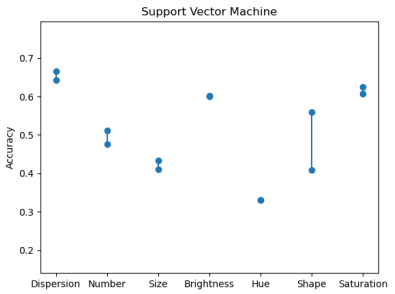
Deriving Emotional Responses from Generated Images Our implementation of the transformer-based approach, utilising the pre-trained Vision Transformer (ViT) model architecture, yielded varying performance levels



(a) Comprehensive Dataset



(b) Individual Pieces



(c) Clustered Pieces

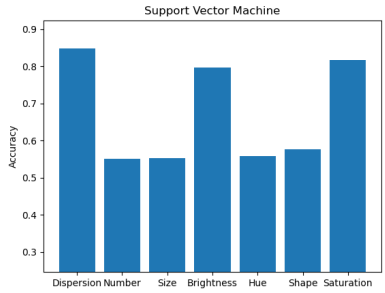
Figure 2.22. SVM model performance for three-category classification of image parameters across different analytical approaches.

across different emotional responses.

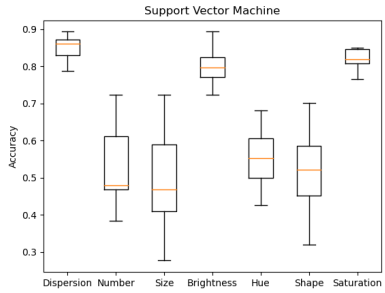
Table 2.18. Performance of Transformer Model in Predicting Emotions from Images

Emotion	MAE	Emotion	MAE	Emotion	MAE
Amazement	0.27	Solemnity	0.35	Tenderness	0.24
Nostalgia	0.25	Calmness	0.33	Power	0.29
Joyful Activation	0.29	Tension	0.29	Sadness	0.25

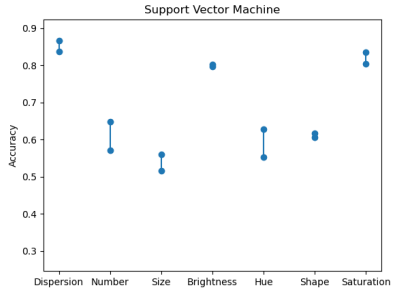
As presented in Table 2.18, the model achieved highest accuracy in predicting Tenderness (MAE 0.24), followed closely by Nostalgia and Sadness (both MAE 0.25). Solemnity (MAE 0.35) and Calmness (MAE 0.33) proved



(a) Comprehensive Dataset



(b) Individual Pieces



(c) Clustered Pieces

Figure 2.23. SVM model performance for binary classification of image parameters across different analytical approaches.

more challenging to predict accurately. The model demonstrated consistent performance (MAE 0.29) for Power, Joyful Activation, and Tension predictions.

Conclusions

Our investigation has examined the intricate relationships between emotion, auditory stimuli, and visual perception, building upon established research in crossmodal perception and emotion-mediated experiences. The findings offer insights into these bidirectional relationships whilst suggesting applications in both theoretical and practical domains.

Forecasting Visual Parameters from Musical Emotional Responses Addressing our first research question, we established the feasibility of estimating visual imagery characteristics from emotional responses to music. Our analytical models demonstrated varying success across different visual parameters:

- Dispersion, Brightness, and Saturation exhibited highest predictability across all analytical approaches (regression, three-category, and binary classification).
- Number, Size, and Hue parameters presented greater challenges in prediction, suggesting more complex or subjective emotional relationships.
- Model performance varied by classification methodology, with binary classification typically achieving highest accuracy levels.

These findings indicate that certain visual elements maintain more consistent associations with emotional responses to music across individuals. This consistency aligns with previous investigations identifying patterns in colour-emotion associations in music perception. The successful prediction of certain visual parameters suggests promising applications in sensory substitution. Our models could potentially support development of systems translating auditory experiences into visual representations, particularly focusing on more predictable aspects such as brightness and colour saturation.

Inferring Emotional Responses from Generated Images Our second research question examined the inverse relationship: deriving emotional responses from generated images. Our transformer-based model revealed:

- Tenderness, Nostalgia, and Sadness demonstrated higher prediction accuracy from visual data.
- Solemnity and Calmness proved more challenging to infer from visual representations alone.

- Power, Joyful Activation, and Tension showed consistent predictability levels, suggesting potential universality in their visual expression.

These results support the hypothesis of common pathways through which emotions influence visual creation across individuals. However, varying predictability across different emotions highlights this relationship's complexity. The capability to infer emotional responses from visual data has significant implications, potentially supporting development of tools for understanding and communicating with individuals who express emotions differently, including neurodiverse populations.

Implications and Future Directions Our investigation contributes to the growing understanding of emotion's role in crossmodal perception by providing quantitative models for both directions of the emotion-visual imagery relationship. The bidirectional nature of our approach offers comprehensive insight into how emotions mediate between auditory and visual experiences. However, our research also identifies areas requiring further investigation:

- Model performance variability across parameters and emotions suggests the need for more nuanced approaches capturing individual emotional-visual association subtleties.

- Prediction challenges for certain visual parameters and emotions indicate potential influence of personal, cultural, or contextual factors not fully captured by current models.

- Ethical considerations regarding AI implementation in emotion simulation and potential manipulation warrant ongoing attention as these technologies develop.

Future research opportunities include exploration of more diverse datasets, including cross-cultural samples, to enhance model generalisability. Additionally, investigating practical applications in therapeutic interventions or personalised entertainment experiences could provide valuable insights

into real-world utility. In conclusion, our investigation demonstrates the feasibility of modelling complex relationships between emotional responses to music and visual imagery, as well as the reverse process of inferring emotional responses from visual data. These findings advance theoretical understanding of crossmodal perception whilst suggesting innovative applications ranging from assistive technologies to creative arts.

2.6 Implications for Human-Computer Interaction

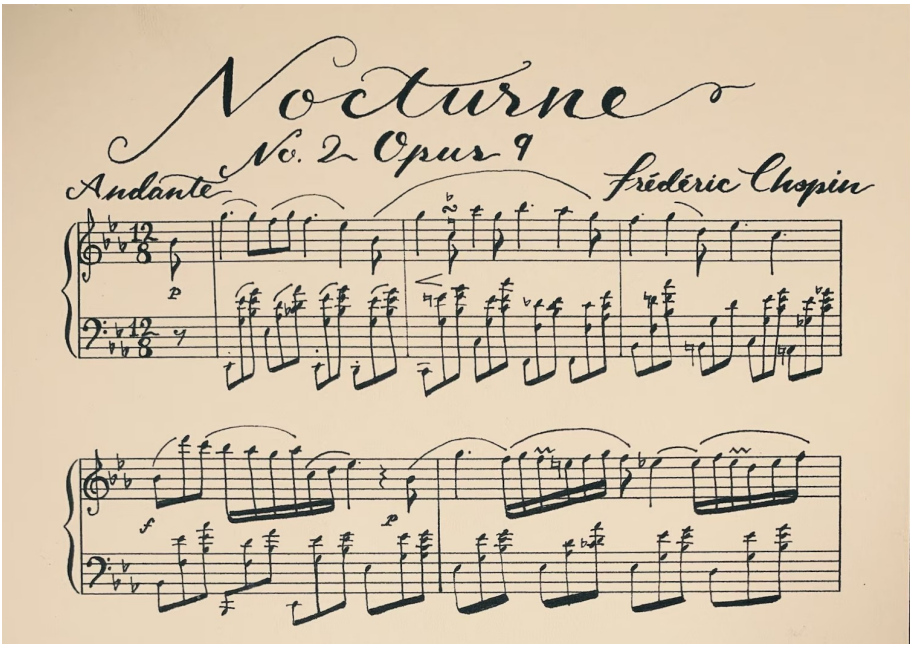
The findings presented in this chapter provide several concrete pathways for advancing human-computer interaction technologies. First, the developmental differences in crossmodal associations identified in Section 2.4 suggest that age-adaptive interfaces could enhance user experience by tailoring audiovisual mappings to specific demographic groups. Our results indicate that while the OC-MO-RT mapping shows consistency across age groups, certain parameter combinations could be optimised differently for children versus adults.

Second, the quantitative models of emotional mediation developed in Section 2.5 offer a framework for evaluating and enhancing emotional congruence in multimodal interfaces. Developers can implement these models as evaluation metrics during the design process, ensuring that audiovisual elements maintain consistent emotional valence. For example, interfaces designed to evoke calmness should consider specific combinations of brightness, saturation, and spatial parameters identified in our analysis. Third, the bidirectional predictive capabilities demonstrated in our research enable novel interaction paradigms where emotional states could be inferred from user-generated visual content or, conversely, where visual elements could be dynamically adjusted based on detected emotional responses to auditory stimuli. Such technologies could support applications ranging from affective computing interfaces to accessibility tools for individuals with sen-

sory impairments. These practical implications extend beyond theoretical contributions, offering concrete methodologies and design guidelines for developing emotionally coherent human-computer interfaces that leverage the fundamental relationships between crossmodal perception and emotional processing

Chapter 3

Affective Computing and Emotion Recognition



Frédéric Chopin, "Nocturne in E-flat major, Op. 9, No. 2" (1830-1832), musical score, first page.

3.1 Physiology of Emotion

3.1.1 Understanding Emotions

Definitions and Distinctions

The topic of emotions is of fundamental importance in human physiology, affecting not only psychological aspects but also a wide range of physiological changes that prepare the organism for specific actions [81]. While emotions are a universal aspect of human experience, providing precise scientific definitions requires careful distinction between related terms such as emotions, feelings, and moods. An emotion can be defined as a complex psychophysiological process that arises automatically and unconsciously in response to significant stimuli. It involves characteristic patterns of physiological activation, specific neural mechanisms, cognitive appraisals, and a tendency toward particular actions [81]. This integrated response pattern manifests through specific behavioural motor patterns and is typically short-lived, lasting from seconds to minutes. Emotions involve well-defined neural circuits and show remarkable universality across different cultures, highlighting their role as fundamental biological responses. Feelings, in contrast, represent the conscious awareness and subjective interpretation of emotional states. While sharing the same underlying psychophysiological foundation, feelings differ from emotions in several key aspects. They can persist for longer periods, from minutes to hours, and may occur without overt behavioural expressions. Unlike the universal nature of basic emotions, feelings are more heavily influenced by cultural and personal factors, showing greater variability across different societies and individuals. This cognitive dimension of feelings involves more complex processing compared to the automatic nature of emotional responses, and may have subtle or no direct physiological correlates. This distinction between emotions and feelings is crucial for understanding both the automatic physiological responses that characterize emotions and the role of consciousness

in emotional experience [81]. While emotions prepare the organism for immediate action through specific physiological changes, feelings provide the conscious framework through which we interpret and make sense of these emotional experiences.

	Emotions	Feelings
Latency	Short	Long
Object	Definite	Indefinite
Corporeity	Essential	Secondary
Voluntary Control	Poor	High

Table 3.1. Main differences between emotions and feelings [81]

A further important distinction needs to be made between emotions and moods. While emotions are rapid responses to specific stimuli or situations, moods represent more prolonged affective states that can persist for hours, days, or even weeks. Moods typically lack a specific trigger and have a more diffuse influence on cognition and behaviour than emotions [81]. Key characteristics that distinguish moods from emotions include:

- **Duration:** While emotions are acute responses lasting seconds to minutes, moods are tonic states that can persist for extended periods
- **Intensity:** Emotions tend to be more intense than moods, involving stronger physiological activation
- **Specificity:** Emotions are typically directed towards specific objects or events, while moods are more generalised and often lack a clear object
- **Behavioural impact:** Emotions tend to interrupt ongoing behaviour and demand immediate attention, while moods operate as background states that colour perception and cognition

The interactions between these different affective phenomena are complex and bidirectional. For instance, repeated emotional experiences of the same type can contribute to the development of a corresponding mood state. Conversely, existing moods can influence the likelihood and intensity of specific emotional responses. This relationship highlights the dynamic nature of affective processes and their importance in human behaviour and adaptation. Understanding these distinctions is crucial not only from a theoretical perspective but also for practical applications in fields such as clinical assessment, psychological research, and the development of artificial systems for emotion recognition. The different temporal dynamics and physiological signatures of emotions, feelings, and moods require different approaches for their measurement and analysis.

Emotion representations

Representations of emotion have been a central focus in psychology, with various theories proposed to capture the complex nature of affective experiences. Two prominent approaches have emerged: the basic emotions theory and the dimensional theory. The basic emotions theory, as proposed by Ekman (1992), suggests that there are a limited number of universally recognised, discrete emotions, such as happiness, sadness, anger, fear, disgust, and surprise [82]. Ekman's work on facial expressions has been influential in supporting this theory. He conducted extensive research on the facial expressions associated with these basic emotions, using photographs of individuals from different cultures to demonstrate the universality of emotional expressions [83, 84].

Ekman's basic emotions theory has been supported by numerous cross-cultural studies, which have shown that individuals from different cultures can reliably identify the facial expressions associated with basic emotions [85, 86, 87]. The basic emotions theory has been widely applied in affective computing and emotion recognition research, where automatic systems are developed to identify emotional states based on facial expressions, vocal cues, or physiological signals [88, 89].

Figure 3.1. Facial expressions of basic emotions according to Ekman. Adapted from [84].

In contrast, dimensional theories argue that emotions are better represented along continuous dimensions. Russell's circumplex model (1980) is one of the most prominent dimensional theories, proposing that emotions can be mapped onto a two-dimensional space consisting of valence (pleasure–displeasure) and arousal (high–low activation) [90].

This model has been supported by empirical studies using self-report measures and physiological data [91, 92]. Some researchers have suggested extending the model to include a third dimension, dominance (control–submission) [93]. Dimensional theories offer a more flexible and comprehensive approach to representing the full range of emotional experiences compared to the discrete categories proposed by the basic emotions theory [94].

However, both the basic emotions theory and dimensional models have

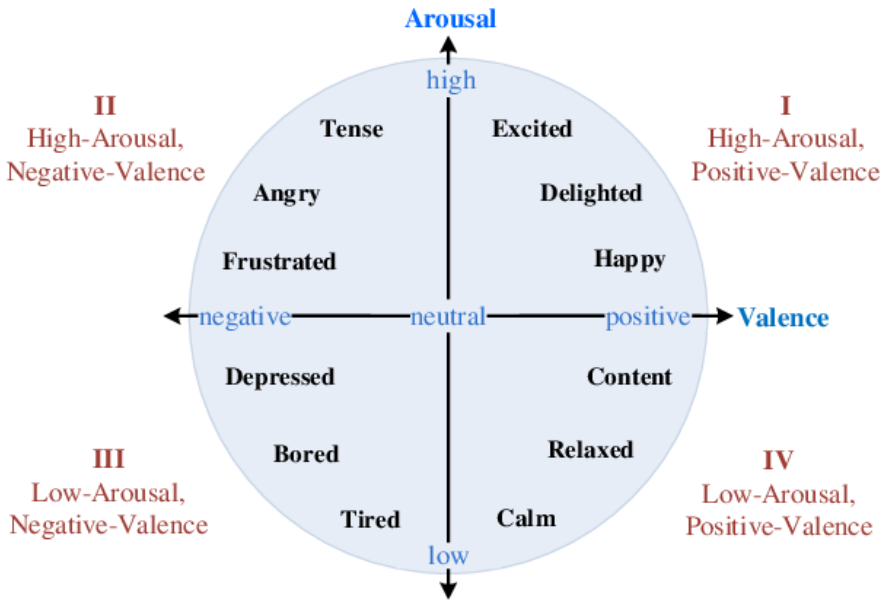


Figure 3.2. The dimensional model of emotions, representing emotions along the valence and arousal dimensions. Adapted from [90].

limitations when it comes to capturing the nuances of emotions induced by specific stimuli, such as music. Music has been shown to evoke a wide range of affective states that may not be fully represented by these general emotion models [95]. Recognising this, Zentner et al. (2008) developed the Geneva Emotional Music Scale (GEMS), a domain-specific emotion representation tailored to music-induced emotions [96]. The GEMS was derived through a series of studies involving factor analysis of listeners' ratings of felt emotions in response to music excerpts. The resulting model consists of nine music-specific emotion factors: wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness [96].

In developing the GEMS, Zentner et al. (2008) demonstrated that music-induced emotions are best captured by a domain-specific model rather than the basic emotions or dimensional models [96]. The GEMS showed better discrimination of musical emotions and accounted for a greater proportion of variance in felt emotions compared to the other models. This highlights

the potential for domain-specific emotion representations to provide a more nuanced understanding of affective experiences in particular contexts, such as music listening [97].

Cognitive Appraisal Theories

Cognitive appraisal theories represent a significant advancement in our understanding of emotional processes, emphasizing the crucial role of cognitive evaluation in emotional experiences. These theories, pioneered by researchers like Lazarus and Folkman, posit that emotions arise not directly from events themselves, but from the individual's interpretation and evaluation of these events [98]. The appraisal process involves two key stages: primary appraisal, where an individual evaluates whether an event is relevant to their well-being, and secondary appraisal, where they assess their coping resources and options [99]. According to Lazarus's theory, the interaction between personal goals and environmental conditions leads to specific emotional responses through cognitive evaluation processes. This evaluation encompasses various dimensions including goal relevance and congruence, the degree of ego involvement, perceived coping potential, and future expectations. The theory suggests that emotional responses are shaped by how individuals interpret events in relation to their personal goals and their perceived ability to cope with the situation [100].

Computational Models of Emotions: The OCC Model

Building upon cognitive appraisal theories, Ortony, Clore, and Collins (1988) developed a computational model of emotions, known as the OCC model, which has become a standard framework for emotion synthesis in artificial systems [101]. This model provides a structured approach to understanding how emotions arise from the evaluation of situations, defining 22 emotion categories based on valenced reactions to events, actions of agents, and aspects of objects.

The OCC model categorizes emotional responses into three main branches

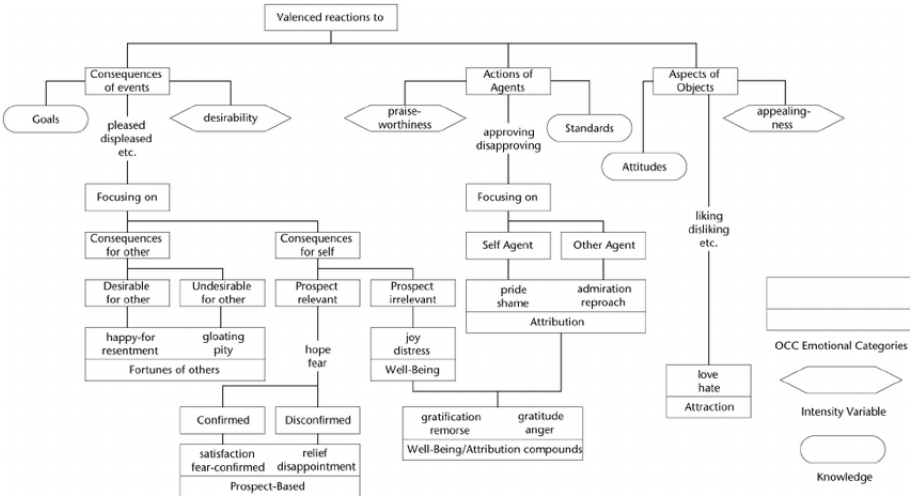


Figure 3.3. The OCC model structure showing the hierarchical organisation of emotional categories based on different types of stimuli evaluation [102]

based on their triggering causes: consequences of events, actions of agents, and aspects of objects. Event-based emotions arise from evaluating outcomes relative to goals, leading to emotions such as joy, distress, happy-for, and pity. Agent-based emotions result from assessing actions against standards, producing emotions like pride, shame, admiration, and reproach. Object-based emotions emerge from evaluating objects based on attitudes, generating emotions such as love and hate. This model has proven particularly valuable in the field of affective computing, providing a systematic framework for implementing emotional responses in artificial agents [102]. The model's structure allows for the quantification of emotional intensity based on the desirability of events relative to goals, the praiseworthiness of actions relative to standards, and the appealingness of objects relative to attitudes. While the OCC model offers a comprehensive framework for emotion generation, it requires several extensions for practical implementation. These include a history function to track emotional experiences over time, mechanisms for emotional state interaction, and personality parameters to ensure consistent behavioural responses [102]. These additions help bridge the gap between theoretical understanding and practical appli-

cation in artificial emotional systems, making the model particularly useful for developing emotionally intelligent artificial agents.

3.1.2 Emotions and Bodily Changes

The relationship between emotions and bodily changes represents a fundamental aspect of understanding emotional experience. This relationship first gained scientific attention through the James-Lange theory, which proposed that emotional feelings are the result of perceiving bodily changes rather than their cause [103, 104]. In James's words, "we feel sorry because we cry, angry because we strike, afraid because we tremble" - a counter-intuitive view that sparked decades of research into emotion-body relationships.

Historical Development of Emotion-Body Theories

The James-Lange theory faced significant challenges from subsequent research, particularly through Walter Cannon's systematic critique [103]. Cannon identified several crucial limitations: total separation of viscera from the central nervous system does not alter emotional behaviour; the same visceral changes occur in different emotional states and even non-emotional conditions; visceral structures are relatively insensitive with limited sensory innervation; and visceral changes are too slow to be the source of emotional feeling. These findings necessitated a more nuanced understanding of how bodily states contribute to emotional experience.

The Somatic Marker Hypothesis

A more comprehensive framework for understanding emotion-body relationships emerged through Damasio's somatic marker hypothesis [105]. This theory proposes that emotional experiences become associated with bodily states through "marker" signals that influence decision-making and behavioural responses at multiple levels. These markers operate through two potential pathways: a "body loop" involving actual physiological changes

and their feedback, and an "as-if body loop" where somatosensory regions directly simulate bodily states without peripheral changes.

The ventromedial prefrontal cortex plays a crucial role in this system by storing and reactivating learned associations between situations and their corresponding somatic states. Evidence for this framework comes from studies of patients with ventromedial prefrontal damage, who show impaired decision-making despite intact general intelligence. These patients exhibit diminished autonomic responses to emotionally charged stimuli and make poor choices in situations that simulate real-life decision-making under uncertainty [105].

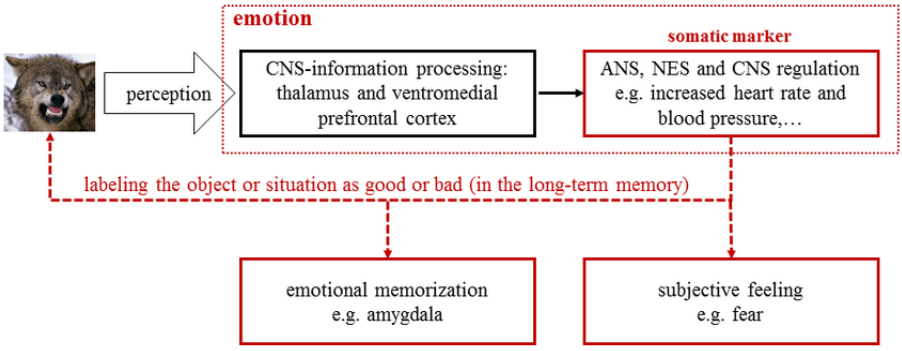


Figure 3.4. Diagram of the process from perception to emotion and subjective feeling. Sensory information is processed by the thalamus and ventromedial prefrontal cortex. Somatic markers guide ANS, NES and CNS regulation, giving rise to emotion. Subjectively experienced emotion generates feeling, e.g. fear. [106]

Modern Understanding of Physiological Response Patterns

Contemporary research recognises a complex bidirectional relationship between emotional experiences and physiological states [81]. The autonomic nervous system orchestrates a range of bodily changes during emotional experiences, including alterations in heart rate, blood pressure, skin conductance, and respiratory patterns. The endocrine system also plays a vital role through hormonal changes, particularly involving the hypothalamic-

pituitary-adrenal axis.

These physiological responses show considerable individual variation influenced by genetic predispositions, previous experiences, cultural background, and current physiological state. However, certain patterns of autonomic and endocrine activation tend to be associated with specific emotional states, though these relationships are more complex than originally proposed by early emotion theories.

Integration of Bodily Changes in Emotional Processing

Modern perspectives emphasize that bodily changes contribute to emotional experience not merely as feedback signals but as integral components of an emotion-cognition network. The somatic marker hypothesis particularly highlights how bodily states help guide decision-making by constraining the decision space and providing rapid, unconscious biasing signals before detailed rational analysis occurs.

This understanding has important implications for both theoretical models of emotion and practical applications in fields such as affective computing and clinical assessment. It suggests that emotional experience emerges from the integration of multiple physiological and neural systems, with bodily changes serving both as components of emotional experience and as guides for adaptive behaviour [105, 81].

The relationship between emotions and bodily changes thus appears more nuanced than either pure cognitive theories or purely physiological approaches would suggest. Instead of being simply a cause or consequence of emotional experience, bodily changes appear to be part of an integrated system that supports both emotional experience and adaptive behaviour regulation. This integration helps explain how emotions can influence decision-making and behaviour both through conscious awareness and through unconscious bodily signals.

3.1.3 Functional Neuroanatomy of Emotions

The neural basis of emotions involves complex interactions between multiple brain regions, forming interconnected circuits that process, regulate, and integrate emotional responses. Modern neuroimaging techniques, particularly functional magnetic resonance imaging (fMRI), have revolutionised our understanding of how the brain processes and regulates emotions [107].

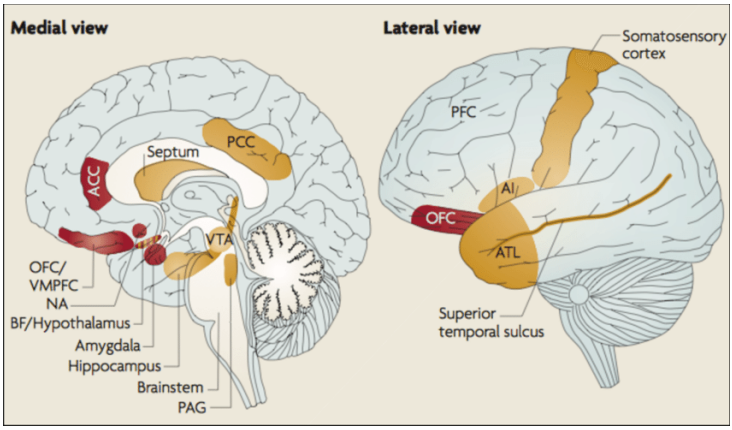


Figure 3.5. The figure illustrates key brain regions associated with emotional processing, divided into core (dark red) and extended (brown) regions. The core regions, which appear frequently in affective neuroscience literature, include subcortical structures such as the amygdala, nucleus accumbens (NA), and hypothalamus, along with cortical areas like the orbitofrontal cortex (OFC), anterior cingulate cortex (ACC), and ventromedial prefrontal cortex (VMPFC). Extended regions include additional subcortical structures (e.g., brainstem, ventral tegmental area (VTA), hippocampus, periaqueductal grey (PAG), septum, and basal forebrain (BF)) and cortical areas (e.g., anterior insula (AI), prefrontal cortex (PFC), anterior temporal lobe (ATL), posterior cingulate cortex (PCC), superior temporal sulcus, and somatosensory cortex). These regions collectively contribute to affective functions, though none are purely emotional in their role.[108]

Neural Circuits of Emotion Processing

Emotional processing relies on distributed networks rather than isolated brain regions. Recent fMRI studies have identified several key intrinsic brain networks, including the Default Mode Network (DMN) and Dorsal Attention Network (DAN), that exhibit overlapping activation patterns during emotional processing [109]. These networks demonstrate a shared neural architecture for processing both motivational and emotional information, highlighting the intricate nature of emotional neural circuits.

Film-based fMRI methodologies have enabled researchers to study emotions in more naturalistic settings, revealing that various emotions activate overlapping cortical and subcortical areas [110]. This suggests that emotions emerge from broader psychological processes involving multiple brain regions rather than being localised to specific areas. Coordinate-based meta-analyses have shown that different types of emotional stimuli engage distinct brain systems while also activating common regions [111].

The Limbic System

The limbic system, traditionally considered the emotional centre of the brain, comprises several interconnected structures that work in concert to process and regulate emotions. At its core, the limbic system includes the amygdala, hippocampus, anterior cingulate cortex, and hypothalamus. These structures form a complex network that integrates sensory information, coordinates autonomic responses, and modulates emotional behaviour [112].

Research has demonstrated that the limbic system's components show specific activation patterns during different emotional states. The anterior cingulate cortex serves as a critical interface between emotional and cognitive processes, while the hypothalamus coordinates autonomic responses to emotional stimuli [113]. The hippocampus plays a crucial role in emotional memory formation, contextualising emotional experiences within our autobiographical memory [114].

Role of the Amygdala

The amygdala serves as a crucial hub in emotional processing, particularly in detecting and responding to emotionally salient stimuli. Meta-analyses of neuroimaging studies have revealed its critical involvement in processing both positive and negative emotional stimuli, with particular sensitivity to facial expressions [115]. Recent research has challenged the traditional view of lateralisation, showing that both left and right amygdala contribute to emotional processing, albeit with distinct temporal dynamics.

Studies investigating emotional responses to music have demonstrated that the amygdala is prominently activated during the perception of dissonant (unpleasant) music, while pleasant music engages regions such as the inferior frontal gyrus and ventral striatum [116]. This highlights the amygdala's role in processing diverse types of emotional stimuli across different sensory modalities.

Prefrontal Cortex Involvement

The prefrontal cortex (PFC) plays a fundamental role in emotional regulation through its various subdivisions. The ventrolateral PFC (VLPFC) has been shown to suppress activity in subcortical regions like the amygdala during emotion regulation tasks, particularly during cognitive reappraisal [117]. The dorsolateral PFC contributes to cognitive control processes, while the medial PFC integrates emotional information with decision-making processes [118].

Research has demonstrated that stronger functional connectivity between the dorsolateral prefrontal cortex and the insula correlates with improved clinical outcomes in depression treatments [114]. This highlights the therapeutic potential of targeting specific prefrontal-limbic connections in treating mood disorders.

Interaction Between Emotion and Cognition

The interaction between emotional and cognitive processes occurs through bidirectional connections between the PFC and limbic structures. These connections enable top-down regulation of emotional responses while allowing emotional information to influence cognitive processes [119]. Neuroimaging studies have shown that cognitive reappraisal of emotions activates prefrontal regions while simultaneously modulating activity in limbic areas, demonstrating the neural basis of cognitive emotional regulation [120].

This interaction is particularly evident in studies of attention and emotion, where emotional stimuli can be processed automatically but require sufficient cognitive resources for optimal processing [121]. Recent research has particularly highlighted the role of the anterior insula in integrating emotional and cognitive information, serving as a crucial hub in the emotional awareness network [122].

Neuroimaging Studies of Emotion

Modern neuroimaging techniques have provided new insights into emotion processing. Machine learning applications to fMRI data have enabled researchers to identify specific neural signatures associated with emotion regulation [123]. These studies have shown that activation patterns in the ventrolateral PFC, dorsomedial PFC, and insula can indicate when individuals are actively modulating their emotional responses to negative stimuli.

Research investigating emotion processing in clinical populations has revealed important insights into the neural basis of emotional disorders. Studies of individuals with alexithymia, for instance, have shown reduced amygdala activation during emotional processing [124], while research on borderline personality disorder has identified distinct patterns of neural activation during emotional tasks [125]. These findings contribute to our understanding of how emotional processing may be altered in various psychological conditions and suggest potential therapeutic targets.

3.2 Methods and Technologies in Affective Computing

3.2.1 Fundamentals of Affective Computing

Affective computing represents a significant paradigm shift in human-computer interaction, moving beyond traditional interface design to systems capable of recognising, interpreting, and responding to human emotions. This section explores the fundamental concepts, methodological approaches, and technical implementations in this rapidly evolving field.

Origins and Evolution of Affective Computing

The field of affective computing emerged from the pioneering work of Rosalind Picard at the MIT Media Laboratory, who first introduced the term in her seminal 1995 paper [126]. Picard's work highlighted the crucial role of emotional intelligence in technology, arguing that for computers to exhibit genuine intelligence and effectively interact with humans, they must have the ability to recognise and respond to human emotions [127]. This marked a significant departure from traditional human-computer interaction paradigms, which primarily focused on efficiency and task completion.

The field has since evolved significantly, with recent bibliometric analyses indicating a substantial increase in research activity, particularly from regions such as China and Western institutions [128]. This growth reflects the increasing recognition of emotion's role in human-computer interaction and the potential applications across various domains, from healthcare to entertainment [129].

Computational Approaches to Emotion Processing

Building upon the theoretical foundations of emotion science, affective computing implements computational frameworks for emotion recognition and response. A significant advancement in this area is the Function-

Component-Representation (FCR) framework proposed by Ma and Yarosh [130], which provides a structured approach to understanding and implementing affective computing systems. The framework can be formally defined as:

$$AC_{system} = f(F_{affect}, C_{affect}, R_{affect}) \tag{3.1}$$

Where:

- F_{affect} represents the function (why compute affect)
- C_{affect} denotes the component (how to compute affect)
- R_{affect} indicates the representation (what affect to compute)

This framework has proven particularly valuable in implementing systems based on established emotion theories, including both discrete and dimensional models [131]. The computational implementation of these models often employs machine learning algorithms, with recent advances in deep learning showing promising results in emotion recognition tasks [132].

Experimental Methods in Affective Computing

The experimental methodology in affective computing encompasses various approaches to emotion elicitation and measurement. Research protocols must carefully balance ecological validity with experimental control, while ensuring reliable and reproducible results.

Emotion Elicitation Methods Affective computing research employs two primary categories of emotion elicitation methods [133]:

- *Active methods* involve direct participant engagement through behavioural manipulation tasks, social interaction scenarios, and interactive virtual environments. These approaches enable the study of emotions in dynamic, interactive contexts that more closely resemble real-world

situations. Performance-based challenges can effectively induce stress or achievement-related emotions, providing valuable data for studying emotional responses in task-oriented scenarios.

- *Passive methods* rely on standardised stimuli to evoke emotional responses. The International Affective Picture System (IAPS) [134] and International Affective Digitized Sound System (IADS) [135] represent well-validated tools for emotion elicitation, providing researchers with standardised stimuli that have been extensively characterized in terms of their emotional impact. Film clips and music excerpts offer the advantage of temporal dynamics in emotion induction while maintaining experimental control.

Experimental Design Protocols Standard experimental protocols in affective computing follow rigorous methodological frameworks. The pre-experimental phase involves careful participant screening, informed consent procedures, and baseline measurements to establish foundational physiological and psychological states. Environmental conditions are strictly controlled to minimize confounding variables, with particular attention paid to factors that might influence emotional responses or sensor measurements.

The stimulus presentation phase implements carefully timed and counterbalanced designs to control for order effects and fatigue. Synchronised multimodal data collection enables researchers to capture various aspects of emotional responses simultaneously, including physiological signals, behavioural expressions, and subjective reports. This multimodal approach provides a more comprehensive understanding of emotional responses while allowing for validation across different measurement modalities.

Assessment Tools and Metrics

The Self-Assessment Manikin (SAM) The Self-Assessment Manikin, developed by Bradley and Lang [136], stands as a cornerstone tool in affective

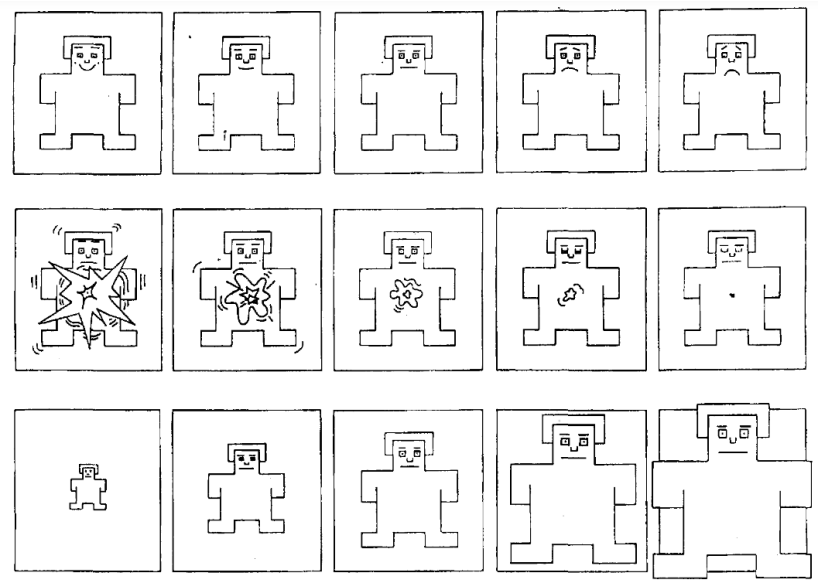


Figure 3.6. The Self-Assessment Manikin (SAM) used to rate the affective dimensions of valence (top panel), arousal (middle panel), and dominance (bottom panel) [136].

computing research, offering a non-verbal pictorial assessment technique that directly measures three fundamental dimensions of emotional experience. This innovative tool presents participants with a series of anthropomorphic figures that represent varying levels of pleasure (valence), arousal, and dominance.

The SAM consists of three distinct scales:

- The valence scale depicts figures ranging from smiling to frowning, representing the pleasantness dimension of emotional experience
- The arousal scale illustrates varying levels of emotional intensity through figures displaying different levels of visceral excitement
- The dominance scale represents the degree of perceived control over the emotional experience, depicted through variations in the figure's size and stance

The tool's nine-point rating system for each dimension provides sufficient granularity to capture nuanced emotional responses while remaining straightforward to administer. Its non-verbal nature makes it particularly valuable for cross-cultural studies and research involving participants with limited verbal abilities. Research has demonstrated strong correlations between SAM ratings and various physiological measures, including heart rate variability, skin conductance, and facial EMG activity [136].

Core Technologies in Affective Computing

Modern affective computing systems typically employ multiple modalities for emotion recognition, integrating various data sources to achieve more robust and accurate emotion detection. The primary modalities include:

Modality	Key Measures	Applications
Facial Expression	Action Units, Expression Intensity	FER Systems
Physiological	HR, HRV, GSR	Stress Detection
Audio	Prosody, Speech Rate	Emotion Recognition
Behavioural	Gestures, Posture	Social Interaction

Table 3.2. Primary modalities in affective computing systems

All the modalities presented in table 3.2 will be explored in more detail in the following paragraph of this section.

Current Challenges and Future Directions

The field of affective computing continues to evolve, facing both technical and methodological challenges. Recent developments in large language models and foundation models have introduced new opportunities and challenges [132]. Technical challenges persist in achieving reliable emotion recognition in real-world conditions, where multiple confounding factors can influence system performance. The integration of multiple

modalities while maintaining real-time processing capabilities remains a significant challenge, particularly in applications requiring immediate emotional assessment and response.

Ethical Considerations

The implementation of affective computing systems raises important ethical considerations that must be carefully addressed [137]. Privacy concerns are paramount, particularly regarding the collection and storage of emotional and physiological data. Researchers and developers must implement robust data protection measures and ensure transparent consent processes that clearly communicate how emotional data will be collected, used, and stored.

The potential for algorithmic bias in emotion recognition systems presents another significant ethical challenge. Systems must be developed and validated across diverse populations to ensure equitable performance regardless of demographic factors. Cultural variations in emotional expression must be considered to avoid misinterpretation or biased results. Additionally, the emotional wellbeing of participants in affective computing research requires careful consideration, with clear protocols for monitoring and managing the intensity of induced emotions and providing appropriate support when needed.

Research integrity in affective computing demands transparent reporting of methodologies and limitations, along with sufficient detail for study replication. As the field continues to advance, establishing and maintaining ethical guidelines becomes increasingly crucial for ensuring responsible development and application of affective computing technologies.

3.2.2 Physiological Measurements in Affective Computing

Physiological measurements provide objective indicators of emotional responses through the activation of the autonomic nervous system (ANS).

These measurements are particularly valuable in affective computing as they offer continuous, non-invasive monitoring of emotional states without requiring conscious input from the participant. This section focuses on the primary physiological measures used in emotion recognition, with particular emphasis on cardiovascular and electrodermal responses.

Cardiovascular Measures

Heart rate (HR) and heart rate variability (HRV) serve as fundamental measures in affective computing research, providing insights into both sympathetic and parasympathetic nervous system activation. Heart rate, measured in beats per minute (BPM), offers a basic indicator of autonomic arousal, while HRV provides more nuanced information about emotional regulation and stress responses [138].

Heart Rate Variability analysis typically employs several time-domain metrics, with the Root Mean Square of Successive Differences (RMSSD) being particularly valuable for assessing emotional responses:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{3.2}$$

where RR_i represents the time interval between successive R-peaks in the ECG signal. RMSSD primarily reflects parasympathetic activity and has shown robust correlations with emotional valence, particularly in response to affective stimuli [140].

Common artifacts in cardiovascular measurements include:

- Motion artifacts from participant movement
- Electrical interference from nearby devices
- Poor electrode contact or placement
- Ectopic beats and other cardiac irregularities

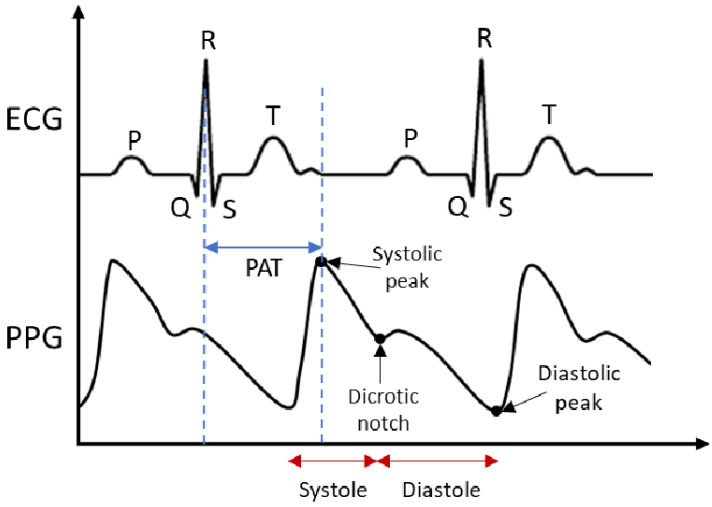


Figure 3.7. The most common sensing technologies employed to monitor cardiovascular activity are the electrocardiogram (ECG) and the photoplethysmogram (PPG). The ECG signal is characterized by specific waveforms (P, Q, R, S, T), with the R-peak being used to calculate R-R intervals for deriving HR and HRV. The PPG signal, reflecting blood volume changes in the microvascular bed, provides inter-beat intervals (IBI) which can similarly be used to calculate HR and HRV. Additional features like the dicrotic notch and systolic and diastolic peaks in the PPG further complement cardiovascular monitoring.HR and HRV [139].

These artifacts require careful preprocessing and typically involve both automated detection algorithms and manual verification.

Electrodermal Activity

Electrodermal activity (EDA), also known as galvanic skin response (GSR), provides a particularly sensitive measure of emotional arousal through changes in skin conductance. EDA reflects purely sympathetic nervous system activity, making it an excellent indicator of psychological arousal and emotional intensity [141].

EDA measurements comprise two main components: Skin Conductance

Level (SCL) and Skin Conductance Response (SCR). SCL refers to the tonic, slowly-changing baseline level, while SCR represents rapid phasic changes in response to stimuli. The interpretation of EDA signals requires consideration of several factors, including individual differences in baseline conductance, environmental conditions such as temperature and humidity, recording site characteristics, and the time course of responses.

Electromyography

Electromyography (EMG) provides valuable insights into emotional states by measuring the electrical activity produced by skeletal muscles. In affective computing, facial EMG is particularly relevant, as it can detect subtle muscle activations associated with emotional expressions, even when these are not visibly apparent [142]. The primary muscle sites for emotional assessment include:

- Corrugator supercilii (associated with frowning and negative emotions)
- Zygomaticus major (involved in smiling and positive emotions)
- Frontalis (related to surprise and attention)

EMG signals require careful preprocessing due to their susceptibility to various artifacts, including cross-talk from adjacent muscles, power line interference, movement artifacts, and electronics noise.

Studies have shown strong correlations between facial EMG activity and emotional valence, with the corrugator showing increased activity during negative emotions and the zygomaticus activating during positive emotions [144].

Blood Pressure and Peripheral Measures

Blood pressure and other peripheral physiological measures provide additional insights into emotional states through cardiovascular system dynamics. Blood pressure, measured as systolic (SBP) and diastolic (DBP)

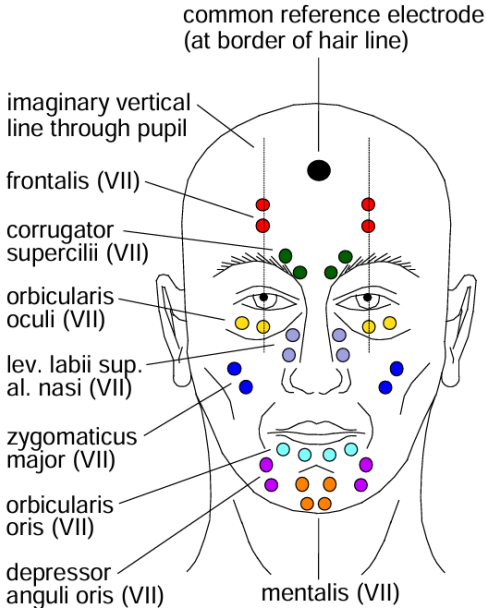


Figure 3.8. Electrode locations for measuring facial EMG activity [143].

pressure, reflects both the direct effects of emotional arousal on the cardiovascular system and longer-term emotional states [145].

Peripheral measures include:

- Peripheral temperature, particularly finger temperature
- Blood volume pulse (BVP)
- Peripheral vascular resistance

These measures are particularly valuable in emotion recognition because they provide continuous, non-invasive monitoring, reflect both short-term emotional reactions and longer-term mood states, and offer complementary information to other cardiovascular measures. However, these measurements face certain challenges, such as sensitivity to movement artifacts, individual variations in baseline measures, environmental influences like temperature and humidity, and time delay in response to emotional stimuli.

Respiratory Measures

While respiratory measures are often secondary to cardiovascular and electrodermal measurements in affective computing, they provide valuable complementary information about emotional states. Respiratory rate, depth, and pattern can indicate emotional arousal and stress levels, though these measures are more susceptible to voluntary control [146].

Correlation with Emotional States

Physiological responses show distinct patterns of correlation with the dimensional model of emotion (valence-arousal). Research using standardized stimuli has revealed consistent relationships:

Measure	Valence	Arousal
Heart Rate	Moderate positive	Strong positive
HRV (RMSSD)	Strong positive	Moderate negative
EDA	Weak/no correlation	Strong positive
EMG Corrugator	Strong negative	Moderate positive
EMG Zygomaticus	Strong positive	Weak positive
Blood Pressure	Weak negative	Strong positive
Peripheral Temperature	Moderate positive	Moderate negative
Respiratory Rate	Weak correlation	Strong positive

Table 3.3. Typical correlations between physiological measures and emotional dimensions

These relationships have been demonstrated across various stimulus types:

Music Stimuli Musical stimuli typically elicit strong physiological responses, with particular effectiveness in modulating both valence and arousal. Research has shown that tempo and mode primarily affect arousal measures

(HR, EDA), while harmonic complexity and consonance correlate more strongly with valence-related HRV measures [147].

Visual Stimuli The International Affective Picture System (IAPS) has been extensively used to validate physiological responses to visual emotional stimuli. EDA responses show particularly robust correlations with image arousal ratings, while HRV measures better reflect the valence dimension [148].

Video Stimuli Video stimuli often elicit the most complex physiological response patterns, likely due to their dynamic nature and multimodal content. Temporal analysis of physiological signals during video viewing has revealed distinct response patterns for different emotional scenarios, with particularly strong EDA responses to suspenseful or surprising content [149].

Common challenges in physiological measurement for affective computing include:

- Individual differences in response patterns
- Context dependency of physiological responses
- Temporal dynamics and response latency
- Integration of multiple physiological channels

These challenges necessitate careful experimental design and appropriate statistical approaches when using physiological measures for emotion recognition in affective computing applications.

3.2.3 Analysis of Non-verbal Behavioural Signals

Non-verbal behavioural signals provide crucial information about emotional states through facial expressions, gestures, and body posture. These signals offer complementary information to physiological measures, often providing more immediate and visually interpretable indicators of emotional responses.

Facial Expression Recognition

Facial Expression Recognition (FER) remains a complex challenge in affective computing, requiring the integration of various contextual components including gender, age, ethnicity, and culture [150, 151]. The development of FER algorithms has focused on leveraging advanced data analysis tools to obtain information about emotional states from facial expressions [152]. These algorithms typically build upon either Ekman's discrete emotion theory [153] or Russell's dimensional theory [90], following a process that begins with face identification, continues with landmark extraction, and concludes with emotion classification [154].

The Facial Action Coding System (FACS), developed by Ekman and Friesen [155], stands as the most well-known and widely used system for analysing facial activity. The automation of FACS represents a significant challenge in the field of behavioural sciences. Despite numerous attempts, achieving fully automated FACS coding remains an unresolved practice. The complexity lies in the association of emotions with facial expressions through the manual, which requires expert human judgment capable of recognising emotional components from frames extracted by AU decoding systems. Nevertheless, various tools have been developed to assist in AU detection and analysis. Notable among these is the Python Facial Expression Analysis Toolbox (Py-FEAT), which represents a significant advancement in automated facial expression analysis [156]. This open-source toolkit enables automated detection of facial landmarks, AU intensity estimation, and emotion classification based on AU configurations, while integrating seamlessly with common deep learning frameworks.

Recent research emphasizes the crucial role of timing in facial expression interpretation, leading to increased focus on algorithms enabling real-time monitoring of facial emotion dynamics [157]. Several pre-made solutions have emerged to facilitate FER applications. The Microsoft Cognitive Service Pack's Emotion API enables face identification and expression analysis, classifying emotions according to Ekman's six basic emotions plus neutrality [158]. However, this API processes only pre-captured images or

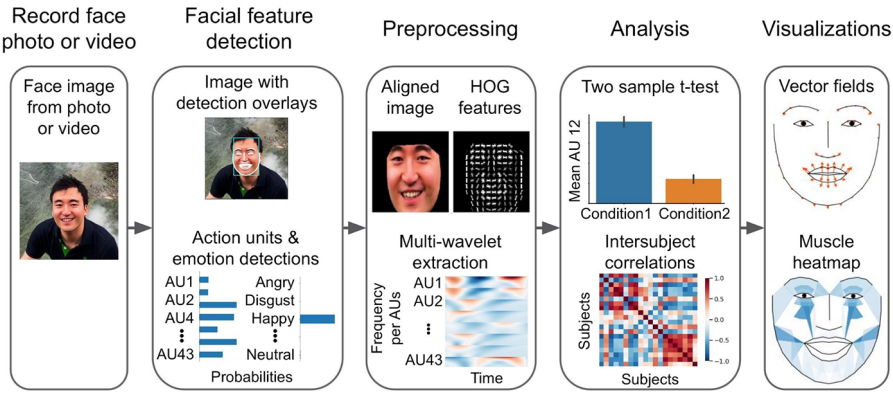


Figure 3.9. Facial expressions analysis pipeline in Py-FEAT [156]. Analysis of facial expressions begins with recording face photos or videos using a recording device such as webcams, camcorders, head mounted cameras, or 360 cameras. After capturing the face, researchers can use Py-Feat to detect facial features such as the location of the face within a rectangular bounding box, the location of key facial landmarks, action units, and emotions, and check the detection results with image overlays and bar graphs.

videos, limiting its real-time implementation capabilities. Affectiva's Affdex technology [159], developed by the MIT Media Lab in 2009, identifies 21 facial expressions and maps them to Ekman's fundamental emotions. The system analyses pixels in 21 facial zones using FACS principles. While Affectiva offers extensive integration options through SDKs for various platforms, research has indicated limitations in its performance for recognising inner emotional states [160, 161].

Face-api.js has emerged as another significant solution, implementing various convolutional neural networks (CNNs) on top of tensorflow.js for face detection and recognition [162]. The library offers comprehensive functionality including face detection, landmark detection, face recognition, and facial expression detection. It employs five pre-trained models: MobilenetV1, TinyFaceDetector, FaceLandmarkModel, FaceLandmark68TinyNet, and FaceRecognitionModel. These models were trained on extensive datasets, with the face detection models utilizing the WIDER FACE dataset (comprising 32,203 images and 393,703 labelled faces), landmark detection mod-

els trained on approximately 35,000 labelled images, and the recognition model trained on over three million images.

The development and evaluation of FER algorithms heavily rely on high-quality labelled datasets [163, 164]. AffectNet stands as one of the most comprehensive datasets, containing approximately one million facial images collected from internet searches using emotion-related keywords in six languages [165]. Of these, about 457,000 images have been manually annotated by experts with seven discrete facial expressions and intensity ratings for valence and arousal. The CK+ dataset [166], expanding upon the original Cohn-Kanade dataset [167], provides both video sequences and static images labeled with seven emotions, including contempt alongside the six basic emotions. The FER-2013 dataset offers 35,887 greyscale images with crowdsourced labels [168], while additional resources like EmoReact, MMI, and RAF-DB provide further options for researchers [169, 164]. Despite these advances, emotion detection in realistic circumstances continues to face significant challenges. These include substantial intra-class variation and low inter-class variation, such as changes in facial position and subtle differences between expressions. The field's continuous expansion demands constant access to high-quality labelled data [170], making dataset selection crucial for specific tasks and available resources. While large-scale datasets can enhance deep learning model performance, smaller datasets may better suit traditional machine learning approaches.

Gesture and Posture Analysis

Body movement analysis enhances emotion recognition through the study of gestures and posture, a field that has gained increasing attention in affective computing research [171]. Research has demonstrated that body movements and postures can effectively communicate emotional states, often providing information complementary to facial expressions [**castellano2007recognising**]. Motion capture methods broadly fall into two categories: marker-based systems using reflective or active markers to track specific body points, and markerless systems employing computer vi-

sion techniques for direct movement tracking.

Feature extraction from gesture and posture data focuses on several key aspects of movement. As outlined by Karg et al. [172], these include kinematic features capturing velocity, acceleration, and jerk, postural features analysing joint angles and body alignment, and quality features assessing movement smoothness and directness. Temporal aspects such as rhythm and synchronization patterns provide additional layers of emotional information. These features find application in various emotional contexts, from detecting stress and anxiety through movement patterns to analysing social interactions and emotional engagement in virtual environments.

Integration with Physiological Measures

The integration of non-verbal behavioural signals with physiological measurements creates a more comprehensive approach to emotion recognition. D'mello and Kory [173] conducted a meta-analysis of multimodal affect detection systems, demonstrating that combining different modalities consistently outperforms unimodal approaches. While physiological measures provide objective indicators of emotional arousal and valence, behavioural signals offer immediate visual cues and capture the social and communicative aspects of emotions. Poria et al. [174] highlight how this complementarity enables cross-validation of emotional state assessments and accounts for both conscious and unconscious aspects of emotional expression. Their review emphasizes the importance of temporal alignment and synchronization between different modalities for effective emotion recognition. Current challenges in non-verbal behavioural analysis include achieving reliable performance in realistic conditions, handling significant intra-class variation, and addressing cultural and individual differences while maintaining real-time processing capabilities. These challenges underscore the importance of continuing research in this field, particularly in developing more robust and adaptable analysis methods.

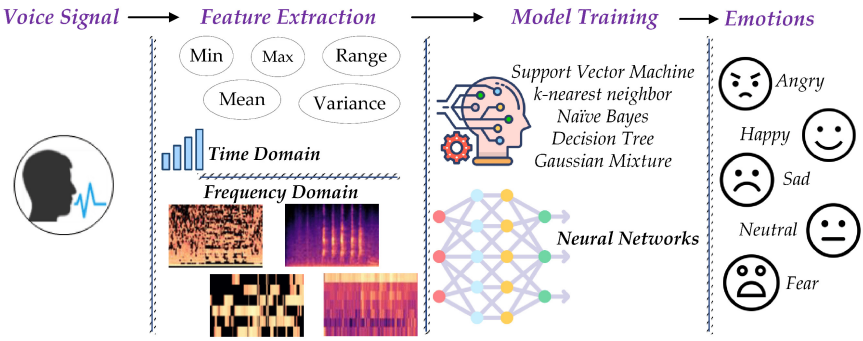


Figure 3.10. An illustration of the SER process using conventional approaches [177].

3.2.4 Audio Technologies in Affective Computing

Speech carries rich emotional information through various acoustic properties, making it a valuable modality for emotion recognition in affective computing. Speech Emotion Recognition (SER) systems analyse multiple aspects of vocal expression to detect and classify emotional states [175]. Speech-based emotion recognition relies on the extraction and analysis of various acoustic features. Prosodic features, including fundamental frequency (pitch), energy, and speaking rate, provide primary indicators of emotional state. These are complemented by spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) and formants, which capture the spectral envelope of speech signals [176]. Voice quality features, including jitter, shimmer, and harmonics-to-noise ratio, offer additional insights into emotional expression through vocal characteristics.

Modern SER systems employ both traditional machine learning approaches and deep learning architectures. While traditional methods rely on carefully engineered feature sets, deep learning approaches can learn relevant features directly from spectrograms or raw audio. Recent advances in deep learning have led to improved performance in real-world conditions, though challenges remain in handling varying acoustic environments and speaker

characteristics [178].

Applications of speech emotion recognition span various domains, from mental health monitoring to customer service analytics. In clinical settings, SER systems can assist in tracking emotional states during therapy sessions or monitoring patients for signs of psychological distress. Commercial applications include call centre analysis for customer satisfaction assessment and virtual assistants capable of adapting their responses based on detected emotional states.

Despite progress in the field, SER systems face ongoing challenges including speaker variability, cultural differences in emotional expression, and the need for robust performance in noisy environments. Integration with other modalities, such as facial expression recognition and physiological measurements, often provides more reliable emotion recognition in practical applications.

3.2.5 Multimodal Integration in Affective Computing

The integration of multiple modalities in affective computing offers a more robust and comprehensive approach to emotion recognition, mirroring the way humans naturally perceive and interpret emotional states. As demonstrated through meta-analysis [173], multimodal systems consistently outperform unimodal approaches in emotion recognition tasks. This section explores the strategies, challenges, and implementations of multimodal integration in affective computing.

Fusion Strategies

Multimodal fusion in affective computing typically employs three main approaches: early fusion (feature-level), late fusion (decision-level), and hybrid fusion [174]. Early fusion combines raw features or feature representations from different modalities before classification, allowing the system to leverage potential correlations between modalities. This approach, however, must address challenges of feature synchronization and dimen-

sionality.

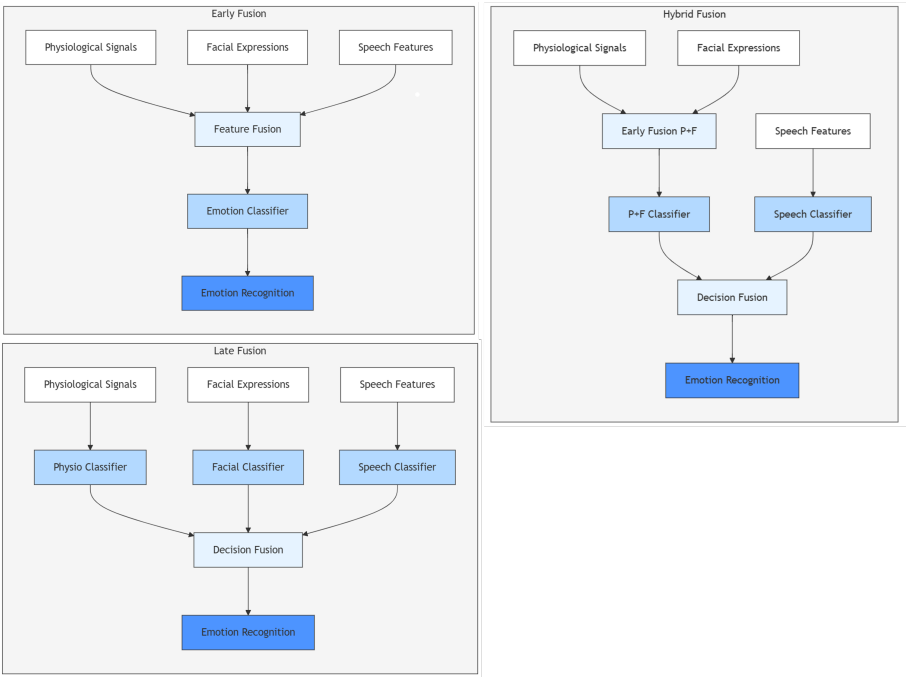


Figure 3.11. Comparison of multimodal fusion architectures for emotion recognition. Early Fusion (top left) combines physiological signals, facial expressions, and speech features at feature level; Late Fusion (bottom) processes each modality through dedicated classifiers before decision-level fusion; and Hybrid Fusion (right) merges physiological and facial data early while processing speech separately. Color intensity indicates processing stage: input features (white), intermediate processing (light blue), and emotion recognition output (dark blue).

Late fusion, conversely, processes each modality independently and combines their decisions or predictions. This approach offers greater flexibility and easier implementation but may miss important cross-modal correlations. Hybrid fusion attempts to capitalize on the advantages of both approaches by combining features at multiple levels of abstraction.

Synchronization and Implementation Approaches

A critical challenge in multimodal integration is the synchronization of different data streams [132]. Each modality operates at different temporal scales: facial expressions may change rapidly, physiological responses show varying latencies, and speech features require sufficient temporal context. Effective integration must address different sampling rates across modalities, varying response latencies and durations, temporal dependencies between modalities, and real-time processing requirements. Successful multimodal systems often employ sophisticated architectures to handle the complexity of integration. Recent approaches utilize deep learning frameworks that can learn optimal fusion strategies directly from data [131]. For example, attention mechanisms have proven effective in weighing the contribution of different modalities based on their reliability and relevance in specific contexts. Several successful implementations demonstrate the effectiveness of multimodal integration, such as combined analysis of facial expressions and speech for more accurate emotion classification, integration of physiological signals with behavioural measures for stress detection, and fusion of multiple modalities for mental health monitoring applications.

Challenges and Future Directions

The implementation of multimodal systems faces several key challenges, including computational complexity and resource requirements, missing or noisy data from individual modalities, calibration and normalization across different signal types, and real-time processing constraints. Additionally, ethical considerations become more prominent in multimodal systems due to the comprehensive nature of the collected data. Privacy concerns, data security, and informed consent require careful attention, particularly when combining multiple data streams that could potentially reveal sensitive personal information [137].

The field of multimodal integration in affective computing continues to evolve, with several promising directions, such as the development of

more efficient fusion architectures, improved methods for handling missing or unreliable modalities, enhanced temporal modelling approaches, and more robust cross-cultural and individual-specific adaptation. As the field advances, the integration of multiple modalities will likely become more seamless and computationally efficient, leading to more reliable and naturalistic emotion recognition systems.

3.3 Research Objectives and Chapter Overview

This chapter addresses fundamental challenges in affective computing through investigations in three interconnected areas exploring different aspects of emotion recognition and response. These studies share common themes in the pursuit of reliable, validated approaches to emotional assessment and interaction, while each focusing on distinct applications and methodological challenges.

The first investigation centres on the validation of Facial Expression Recognition (FER) algorithms, addressing the critical need for standardized evaluation methodologies in emotion recognition systems. This work, published in Sensors in 2023 [179], introduces a universal validation protocol implemented through a web-based framework, enabling systematic assessment of FER algorithms' performance. Additionally, this research contributes to the field through the development of the FeelPix database, providing carefully labelled facial expression data that captures subjects' self-reported emotional states alongside their facial expressions.

The second study explores the integration of emotional awareness in sensory substitution devices through the AFFECT-SENSE system. This novel approach combines real-time physiological measurements with facial expression analysis to create emotionally congruent visual-to-auditory transformations. The research addresses the crucial challenge of maintaining emotional coherence in sensory substitution, potentially enhancing user engagement and device adoption through emotionally appropriate audio feedback. A rationale of this work was presented during the 12th International Conference on Affective Computing & Intelligent Interaction in Glasgow, UK, in September 2024 [180].

The third investigation examines the application of affective computing in music therapy settings, focusing on the development and validation of technology-enhanced therapeutic interventions. This work explores how emotional recognition and response systems can support clinical music therapy practices, contributing to the understanding of music's role in emotional

regulation and expression within therapeutic contexts.

These research areas converge in their emphasis on developing and validating emotion-aware technologies that can be effectively implemented in real-world applications. The subsequent sections present detailed investigations into each of these areas, including theoretical foundations, experimental methodologies, and validation studies. Through these investigations, this chapter contributes to the advancement of affective computing by addressing fundamental challenges in emotion recognition, response generation, and practical application.

3.4 Methodological Approach and Justification

The selection of specific methodologies for emotion recognition and analysis in this thesis was guided by several key considerations, balancing theoretical foundations, practical constraints, and application requirements. This section outlines the rationale behind the methodological choices made across the different modalities of affective computing presented in this chapter.

3.4.1 Rationale for Multimodal Approach

While numerous unimodal approaches exist for emotion recognition, this research deliberately employed a multimodal framework that integrates physiological, behavioural, and audio technologies. This decision was driven by several critical factors:

- **Complementary Information:** Different modalities capture distinct aspects of emotional expression that may not be accessible through a single channel. Physiological measures provide objective indicators of autonomic arousal, facial expressions reveal social communicative aspects of emotion, and audio features capture the temporal dynamics of emotional vocalization [173].

- **Increased Robustness:** Single-modality approaches are vulnerable to specific limitations—physiological signals may be contaminated by motion artifacts, facial expressions can be deliberately suppressed, and audio may be compromised by environmental noise. Multimodality provides redundancy that enhances system resilience in real-world conditions [174].

- **Cross-Validation:** Multiple channels enable cross-validation of emotional assessments, reducing the likelihood of false detections and increasing overall confidence in the system's evaluations [132].

- **Alignment with Natural Perception:** Human emotion perception naturally integrates multiple sensory channels. A multimodal approach more closely emulates this natural processing, potentially leading to more intuitive and human-interpretable systems [181].

Alternative approaches considered included deep neural network architectures specialized for emotion recognition, such as EmoNet [182] and EmotioNet [183]. While these systems demonstrate impressive performance on benchmark datasets, they were not selected due to their substantial computational requirements, limited interpretability, and challenges in real-time implementation across diverse healthcare settings—considerations that were particularly important for the music therapy applications presented in this chapter.

3.4.2 Selection Criteria for Physiological Measurements

Among the wide array of physiological measures available, our research prioritized cardiovascular (HR, HRV) and electrodermal (EDA) activity for several key reasons:

- **Non-Invasiveness:** These measures can be collected using wearable sensors with minimal discomfort to the participant, aligning with our core objective of developing unobtrusive monitoring solutions.

- **Established Relationships with Emotional Dimensions:** Both measures have well-documented relationships with the valence-arousal dimensional model of emotion. HRV has demonstrated consistent correlation with emotional valence, while EDA shows robust relationships with arousal [140].

- **Temporal Responsiveness:** These physiological measures demonstrate sufficiently rapid response characteristics to capture the dynamic nature of emotional changes, while also showing enough stability to provide reliable measurements over extended periods [141].

- **Implementation Feasibility:** The selected measures can be reliably monitored using commercially available, relatively affordable sensor technologies, enhancing the translational potential of the research.

Alternative physiological measures considered included neural measurements (EEG, fNIRS) and respiratory monitoring. Despite their potential for providing additional emotional information, these were not incorporated due to greater susceptibility to movement artifacts (EEG), more complex interpretation requirements (fNIRS), and more intrusive measurement apparatus (respiratory belts). These limitations would have compromised the non-invasive nature of our approach and complicated implementation in ecological settings like those found in the music therapy applications.

3.4.3 Selection Rationale for Facial Expression Analysis

For facial expression analysis, our research employed the Python Facial Expression Analysis Toolbox (Py-FEAT) over alternatives such as OpenFace, Microsoft Cognitive Services, and Affectiva. This selection was based on several considerations:

- **Open-Source Framework:** Py-FEAT's open-source nature allowed for greater customization and integration with other system components, enhancing the reproducibility of our research [156].

- **Comprehensive Analysis Pipeline:** The framework provides a complete pipeline from facial detection to emotion classification based on Facial Action Units (AUs), allowing for both categorical emotion recognition and AU-level analysis [156].
- **Research Orientation:** Unlike commercial solutions that often operate as "black boxes," Py-FEAT was developed specifically for research applications, providing greater transparency in its operation and allowing for more nuanced interpretation of results.
- **Integration Capabilities:** The framework's Python implementation facilitated seamless integration with our other analytical tools and data processing pipelines.

While commercial systems like Affectiva's Affdex may offer more robust performance in certain contexts, their closed nature and subscription requirements limited their suitability for our research purposes. Additionally, OpenFace, while powerful for facial landmark detection, provides less direct support for emotional classification compared to Py-FEAT.

3.5 Universal Validation Protocol for FER algorithms and FeelPix Database

Objectives

From the analysis of the state-of-the-art it emerged that the evaluation and verification of FER algorithms are challenging yet critical. For this reason, the first goal of this work is to devise a universal validation methodology, applicable to any algorithm, aimed at evaluating the performance of Facial Expression Recognition (FER) algorithms. To do so, a web framework was designed, capable of incorporating different algorithms for their investigation. Within this framework, a specific algorithm was tested on healthy subjects, recording their expressions, and comparing the results with their declared emotional state. To elicit an emotional response, the framework presents the user with highly emotive images carefully selected from a specific database. Due to its wide use in various research projects and application domains [184, 185] even if lacking a thorough accompanying documentation, in this project we chose to validate the FER algorithm implemented in the JavaScript library face-api.js. From reverse engineering the code, we observed that the face-api.js face expression recognition model utilizes depthwise separable convolutions and densely connected blocks in its architecture based on ResNet, a popular model for many image recognition tasks due to its ability to train deep networks using skip connections or shortcuts. The model training dataset included diverse images from public datasets and the internet.

From the analysis of the literature, the necessity for annotated databases to develop FER algorithms has become apparent. Presently, existing datasets mainly consist of images, requiring processing through Convolutional Neural Networks (CNNs) or preprocessing to extract facial landmark coordinates for use in machine learning classifiers. Another crucial aspect pertains to the labelling of these datasets, which has been carried out by developers rather than by the subjects themselves, the very individuals from whom the

facial expressions were captured. This phenomenon introduces a significant gap in the data, as it fails to fully capture the genuine emotional responses of the test participants.

Hence, the secondary objective of this study is to collect the subjects' facial expressions' data during the testing phase, thereby crafting an accurate and comprehensive dataset of labelled information. This dataset will serve as a valuable resource for training versatile FER algorithms that can discern users' facial expressions with utmost precision, leveraging facial landmarks as the foundation for analysis.

As a final objective, we have planned to develop a FER algorithm that, leveraging machine learning methodologies, allows for the verification of the reliability of the database constructed, and can be used as an effective methodology for the classification of facial expressions based on facial landmarks coordinates.

Therefore, this research aims to provide a significant contribution to the understanding and development of techniques for facial expression recognition. This work consists of the presentation of a validation method that allows for verifying the reliability of any FER algorithm implementable in web applications. Additionally, this study includes the creation and release of a labelled database, named FeelPix, which can be used to train and test any FER algorithm that recognises facial expressions based on facial landmarks coordinates. Furthermore, the research involves the development of a computationally lightweight yet sufficiently performant algorithm, enabling the evaluation of the effectiveness of the released database by recognising facial expressions based on the data included within it.

Materials and Methods

Experimental Protocol **Experimental setup.** A single subject at a time was tested using a web application in an environment without external sources of lighting, in order to make the results comparable without influencing the user's expressiveness. In order to ensure uniformity of results among all participants, the same computer was used, a Lenovo Essential

V15-IIL notebook with the following characteristics:

- Intel Core i7-1065g7 processor;
- 8 GB Ram;
- Hd 512 GB SSD;
- Display 15.6″;
- Windows 11;
- 0.3 MP front camera resolution.

Google Chrome and Web Server for Chrome applications were used to run the web application in full-screen mode after preloading it from a specific folder.

The language used for the instructions, the evaluation system, and to communicate with participants was Italian.

Protocol. Each participant was provided the appropriate instructions regarding the test procedure before its execution. In addition, the graphical interface informed users about the progress of their tests and subsequent activities. Participants were also asked to timely report any discomfort they felt so that the experiment could be discontinued. A pool of 70 images was identified prior to the experimentation within a specific dataset of highly emotive images, in order to derive the images to be presented to the subjects. Each participant was presented with a randomized subset of 35 images in order to display five images for each of the seven considered emotions (i.e., the six Ekman basic emotions and the neutral emotional status), in a non-consecutive manner. This allowed for variation in the set and order of images presented to each subject, ensuring the generalization of the results thus enhancing the internal validity of the study.

The decision to present participants with a subset of 35 images was grounded in a meticulous assessment of test timing. During the study's design phase, a series of pilot tests were conducted to determine the average time a participant would take to complete the emotional test using varying

quantities of images. It was revealed that with a selection of 35 images, we were able to maintain the total test duration within 10 min. We deemed it essential to prevent participants from experiencing boredom or discomfort due to the test's length, as this situation could adversely affect their emotional expressiveness. Additionally, an excessively long test duration could have led to increased variability in emotional responses, making data interpretation more challenging.

Concurrently, specific guidelines were followed in presenting images to the participants: first of all, five photos were displayed in a non-consecutive order for each emotion (i.e., neutrality, happiness, anger, fear, sorrow, disgust, and surprise). Each image was shown in full-screen mode for 3 s, and the face-api.js library was used to identify emotions and facial cruxes (i.e., landmarks). After this step, the image was reduced and participants were able to see the emotion rating scale. During the decision-making phase, each participant used a seven-button rating scale to describe the emotional state aroused. Each button is labeled with the associated emotion, supported by the corresponding emoticon. In order to avoid affecting the results, the detection of emotions was switched off throughout this decision-making phase. Furthermore, no time restriction was set on the participants in order to give them as much freedom as possible in their choice of emotional state. In fact, each participant was given the choice to move to the next image by pressing a specific button only when he or she thought it was suitable. However, this button was designed in a way that it could only be used after choosing at least one emotional state. Additionally, the web application allowed users to pick multiple emotions for each image simultaneously and change their minds until they had moved on to the next image. Finally, a break time was implemented between each image, during which the subject was informed about the test's progress.

The task was completed in 5–9 min without applying any time restriction.

Participants. A total of 31 healthy subjects (13 females and 18 males) aged between 20 and 69 years participated in the test phase; the age dis-

tribution of the participants can be observed in the graph presented in Figure 3.12.

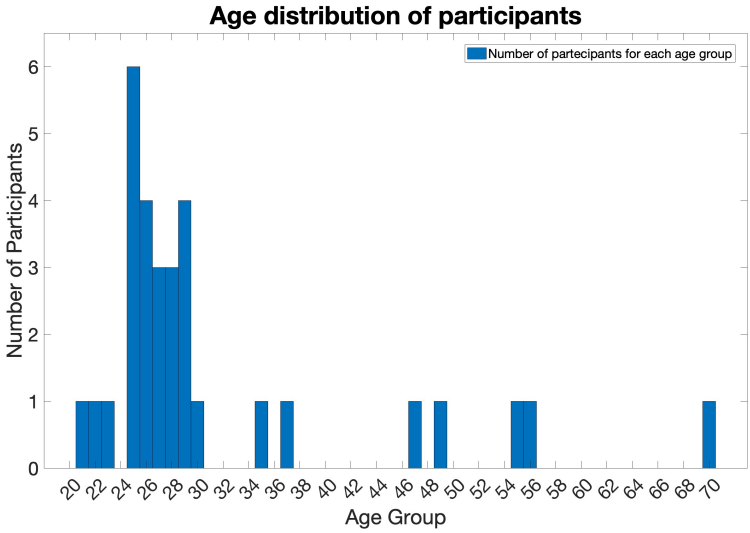


Figure 3.12. Age distribution among individuals involved in the experimental validation process of the Facial Expression Recognition algorithm under investigation.

The data collected for each participant were: first name, surname, date of birth, gender and whether or not they would wear glasses during the test. All subjects voluntarily participated in the study, and no form of compensation was provided. The study was approved by the local ethical committee.

Selection of Highly Emotive Images The selection of appropriate stimuli to consistently induce emotional states played a pivotal role in the experimental protocol. Established stimulus sets are crucial data sources, providing researchers with the means to control and manipulate experimental conditions effectively. At present, numerous well-established, highly emotive stimulus sets are readily accessible. These stimulus sets, carefully selected to evoke emotions consistently, undergo a rigorous standardization process aligned with either the discrete emotion theory and/or the dimensional theory of emotions. As a result of these comprehensive standardiza-

tion efforts, a robust array of stimulus sets has been crafted to harmonize with established theoretical frameworks in the field of emotion.

In this study, it was decided to provide individuals with static visual stimuli, i.e., images, as suggested by specialized literature [186, 187]. We preferred this stimuli over dynamic stimuli such as video or audio tracks, since the aim was to capture instantaneous emotions rather than the temporal evolution of emotional states whereas the use of dynamic stimuli such as video or audio tracks may have introduced confounding factors related to the length and complexity of the stimuli, potentially masking or altering the immediate emotional responses being investigated. Finally, the static images allowed for fine control over the emotional content presented to the participants.

Three different image databases were taken into account [188]: the International Affective Picture System (IAPS), the Open Affective Standardized Image Set (OASIS), and Nencki Affective Picture System (NAPS).

The first one, the IAPS database, developed by Lang et al. [134], provides normative ratings of emotions (pleasure, arousal, dominance) for a series of color images that constitute a set of normative emotional stimuli for experimental investigations of emotion and attention [189]. In this database, each item is accompanied by a set of norms (mean and standard deviation) along three dimensions: arousal (physiological activation evoked by the image), valence (pleasantness and pleasure) and dominance (the degree of control of the emotional state by the subject). More recently, the IAPS has also been standardized according to the discrete emotion theory. The usage of this database is granted by its owner upon request, with the condition that it is solely used for research purposes. The authors of this work applied for access to the database, but the approval for access has not yet been given.

The OASIS [190], is an open-access online stimulus dataset containing 900 color images depicting a broad spectrum of themes, including humans, animals, objects, and scenes, together with normative ratings on two affective dimensions: valence (i.e., the degree of positive or negative affective

response that the image evokes) and arousal (i.e., the intensity of the affective response that the image evokes). The OASIS images were collected from online sources and the ratings of valence and arousal, expressed according to the dimensional theory of emotions, were obtained through an online study. The main advantage of this database lies in its free use for research purposes.

The last database considered is the Nencki Affective Picture System (NAPS) [191], which today is the richest database of visual stimuli with extensive semantic information and a complete set of related normative evaluations. It consists of 1356 realistic photos that have been split into five categories: people, faces, animals, objects, and landscapes. Each image is characterized by a series of emotive ratings that were collected through a test phase on 204 subjects. In particular, the analysis was carried out using the Self-Assessment Manikin (SAM) [136], i.e., a non-verbal pictorial evaluation technique which produces results based on the dimensional theory of emotions. Specifically, using the SAM, each participant expressed the emotional state elicited by the image in terms of valence, arousal, and approach-avoidance. In addition to emotional information, the NAPS provides physical attributes about the images such as brightness, contrast, and entropy. The NAPS was followed by three expansions, including the NAPS Basic Emotions (NAPS BE) [192], which was used in this study.

We decided to use the NAPS BE database rather than the OASIS database, as the algorithm under investigation in the experimental phase yields results based on the discrete theory of emotions to assess the emotional state of the subjects.

This database is a subset of 510 NAPS images that provides classifications based on both the discrete emotion theory and the dimensional theory of emotions. In particular, the NAPS BE includes images that belong into the same five categories as the original database split into the following quantities: 98 animals, 161 faces, 49 landscapes, 102 objects, and 100 people. In order to provide the characterization in discrete terms of this subset of images, a test phase with 124 subjects was developed by the database au-

thors.

In this study, a carefully selected subset of NAPS BE images was used to reduce the number of considered elements and enhance the effectiveness of the experimentation. A selection method was developed to ensure impartiality and prevent our biases from affecting the screening of the images. This method, implemented in Matlab R2022a, utilizes the information provided by the information table associated with the database. Using this approach, we were able to isolate 70 images and divide them into seven groups, each corresponding to one of the seven emotional states considered by the discrete emotion theory. In particular, because images frequently had several labels associated with them, it was not possible to select and categorize them using only discrete labels. For this reason, for each image we took into consideration the intensity values associated with each of Ekman's six basic emotions, provided in range $[1, 7]$, and valence and arousal values, provided in range $[1, 9]$.

The valence and arousal values were subjected to analysis using the k-means clustering algorithm, aiming to identify emotional clusters for grouping the images within the database. The reason for choosing this algorithm is that it always converges and has a very light computational overhead. During the process, a number of clusters equal to three was set in order to obtain three distinct groups of images, each of which could be associated with one of the discrete macro-categories of emotions introduced by Ferré et al. and Kissler [193, 194], namely in Positive, Negative and Neutral. At the same time, another relevant finding was the link, identified by the same authors, between the labels of these macro-categories and mean valence values, such as 2, 5 and 7. In fact, based on this information, each of the three clusters was associated with the corresponding label according to the following logic:

- Negative label to the cluster of images with valence values predominantly in the range $[1, 4]$;

- Neutral label to the cluster of images with valence values predomi-

nantly in the range [4, 6];

- Positive label to the cluster of images with valence values predominantly in the range [6, 9].

Once these three clusters were identified, specific pools of images had to be extracted in order to associate them with the seven emotional states. To do so, the discrete information in the database, namely the discrete labels of each basic emotion and the mean intensity values associated with each image, was utilized.

Firstly, to locate in which of the three clusters to search for the specific images for each discrete label, it was necessary to establish a relationship between each discrete emotion and the clusters previously created. Several methodologies for remapping discrete emotions in the dimensional space are reported in literature. The philosopher Russell J.A., known for developing the circumplex model of emotions, provided a characterization of discrete emotions in dimensional terms [195, 196].

Moreover, statistical analysis methodologies [197, 198] were employed to establish a connection between the two ways of interpreting emotions, such as associating discrete emotion labels with corresponding valence and arousal value pairs. For instance, the emotion label "happiness" may be linked to a specific combination of valence and arousal values. We combined various studies on literature to evaluate how the discrete emotions were distributed in the dimensional space and identify one or more clusters to search for images for each specific discrete emotion, as shown in Table 3.4.

Considering Table 3.4, a noteworthy distinction arises in the treatment of images associated with the emotion of 'surprise' compared to those linked to other emotional states. Specifically, whereas images correlated with other emotional states were confined to particular clusters, images linked to the 'surprise' emotion were deliberately examined across all clusters. This strategic decision is rooted in the inherent universality of the 'surprise' emotion. Research examining the dualistic nature of discrete emo-

Discrete Emotions	Valence [1, 9]	Arousal [1, 9]	Cluster
Anger	3.2	7.7	Negative
Anger	3.2	7.7	Negative
Disgust	2.6	6.4	Negative
Fear	2.4	7.4	Negative
Happiness	8.0	6.9	Positive
Neutrality	5.0	3.0	Neutral
Sadness	2.4	6.1	Negative
Surprise	6.6	7.7	All clusters

Table 3.4. Results of the mapping between discrete and dimensional emotions, with cluster identification for each label.

tions [199, 200], underscores the necessity of an accurate conceptualization of the 'surprise' emotion that embraces both its positive and negative components. In contrast to the categorical classifications typical of other emotions, the complexity of 'surprise' arises from the simultaneous interweaving of these opposing dimensions. This comprehensive perspective elucidates the rationale underlying the distinctive approach adopted when seeking images associated with the 'surprise' emotion, as illustrated in Table 3.4.

The images associated with each specific emotion were searched in the corresponding clusters, considering only those elements that the database developers had labeled with at least the same discrete label as the emotion under consideration. For each emotion, the isolated elements were then rearranged in descending order according to the intensity associated with the emotion in question, in order to select the first ten images to be used for composing the database used in the test phase.

However, it should be noted that:

- Before searching the images for each emotion from the specific clusters, a pre-screening procedure was carried out on the images. In fact,

to avoid excessively shocking the subjects' sensitivity, images that had been assigned an intensity value associated with the emotion "disgust" greater than 4 by the database developers were not considered.

- The neutral emotional state is the only one for which it was not possible to isolate the ten images considering only the labels and intensity values, as the latter information is not provided in the database information table. Therefore, an alternative procedure was implemented:

    1. It was decided to use the discrete labels and arousal values provided.
    2. Images belonging to the neutral cluster, for which the database developers had assigned all six basic emotion labels, were considered.
    3. These images were then reordered according to arousal values, using an increasing sorting order.
    4. Finally, the top ten images with the lowest arousal values were isolated.

In the end, the application of this procedure allowed us to identify ten images for each of the discrete emotions exploited by the face-api.js library (i.e., neutrality, happiness, surprise, fear, sadness, anger, and disgust), resulting in a total of 70 images to be used in the experimental protocol.

Web Application The test was administered through a web app, where both the front-end and back-end were developed using JavaScript. Specifically, the front-end was developed using p5.js, while the back-end was developed using the JavaScript runtime environment Node.js. Consequently, the data obtained throughout the testing phase was saved in the cloud database hosted by the Google Firebase platform.

The decision to utilize p5.js, an open-source Javascript library that provides a comprehensive set of graphical tools, allowed for the creation of a visually engaging and interactive interface. This enhanced the overall user

experience and facilitated the seamless execution of the research. In fact, before administering the test to the subjects, an evaluation of the design and acceptability of the developed application was conducted to validate its efficacy. In this validation test, laboratory students participated, and they provided positive feedback on the interface, describing it as user-friendly, intuitive, and visually appealing.

Specifically, the interface comprises a sequence of user-friendly pages, ensuring inclusivity, explanation, and ease of use throughout the clinical trial experience. The first page provides a brief summary of the activity to be carried out and an input section where the assigned users' identification code is entered, followed by a specific button to start the test. Once the test is started, the emotional image is displayed full screen for 3 s. The interface was developed in order to enable the FER algorithm under investigation (i.e., face-api.js) during this time interval. In this way, it is possible to detect the emotional states expressed through facial expressions for each image and the crucial points of the face for each subject. After this interval, the page for selecting emotional states appears automatically. This page features a smaller version of the image and seven buttons that correspond to the seven considered emotions, allowing the user to select the emotions, one or more, experienced while viewing the image. The transition to the next image occurs without time constraints through a specific button, enabled only after the user has selected at least one emotional state. A 3 s break page has been added before viewing the next image to update participants on the progress of the test. Upon the expiration of this time, automatic redirection to the next page takes place. The choice to set these intervals to 3 s was made to ensure contextual consistency with the experimental protocol adopted by the developers of the highly emotive image content database used in the study [192].

The same sequence of procedures described for the first image is repeated until all 35 images have been viewed and evaluated. The test concludes with a final page that informs the user of the test's completion and expresses gratitude for their participation.

In conclusion, based on the feedback provided by the test subjects, it can be asserted that the interface is easily understandable, intuitive, and visually appealing. Figure 3.13 shows some of the pages of the user interface just described.

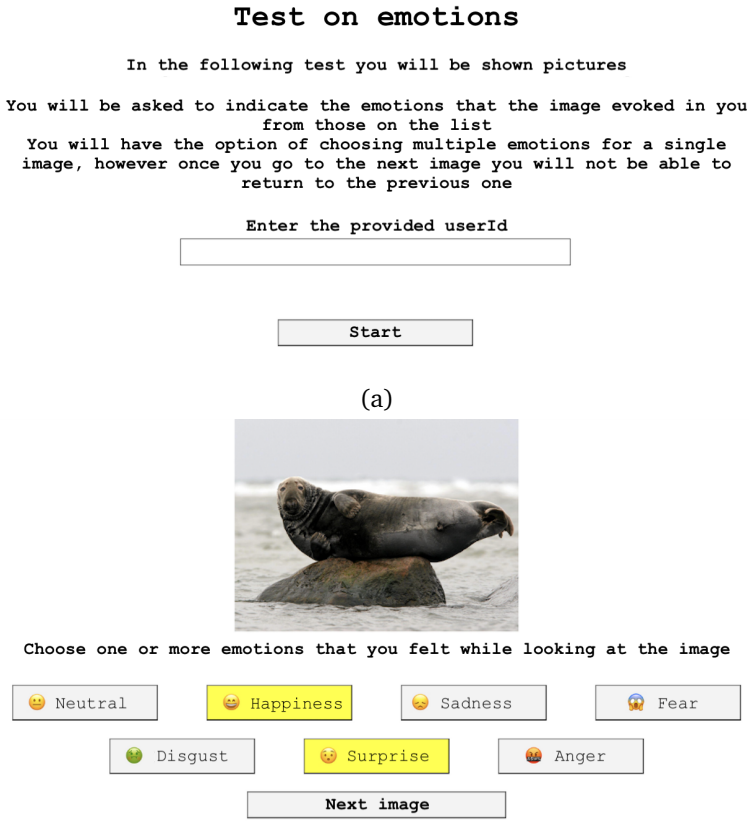


Figure 3.13. Examples of the graphical interface pages designed for the experimental protocol: (**a**) example of the starting page; (**b**) example of the selection page. (Translated from the original Italian version).

FeelPix Dataset During the validation process, specific participant information was extracted and utilized for the development of a labelled dataset.

For the 31 healthy participants (13 females and 18 males) ranging in age from 20 to 69, as depicted in the distribution chart in Figure 3.12, two-dimensional coordinates of the 68 facial landmarks were extracted. It was chosen to isolate the information associated with landmarks as they represent reference points that allow for identifying and tracking particular facial features, including facial expressions [201, 202]. Additionally, emotional states, specified by each participant for every presented image, were collected. These data served as the foundation for constructing the labeled dataset.

This allowed us to have, for each dataset sample, a set of 68 x, y coordinates, accompanied by labeling determined by the ground truth. The ground truth was derived from the emotions specified by users during the experimental phase, resulting in seven labels that characterize each dataset sample. Each of the seven ground truth labels corresponds to one of the seven emotions that the user could indicate in response to the emotional state induced by viewing the image. Binary classification was utilized to translate user selections into discrete data, where a value of 1 was assigned to the chosen emotion labels, while all unselected emotions were labeled with a value of 0. This procedure was applied to each sample, ensuring a consistent mapping between sets of coordinates and a sequence of 1 s and 0 s corresponding to the sample's labels, resulting in a meticulously labeled dataset.

Once the ground truth for the dataset was constructed, to make the raw data detected compatible with the most common standards, a processing procedure was implemented. As a first step, all landmark coordinates were normalized to reduce the impact of different subjects' face positions in the webcam's field of view and their different physiognomy. To achieve this, four specific anatomical points of the face, whose position is not influenced by facial expression, were identified:

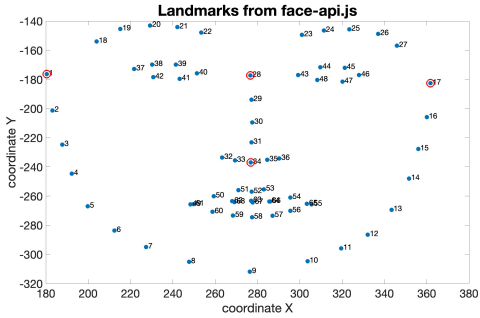
- Left meningeal
- Right meningeal

- Nasal centre
- Subnasal centre

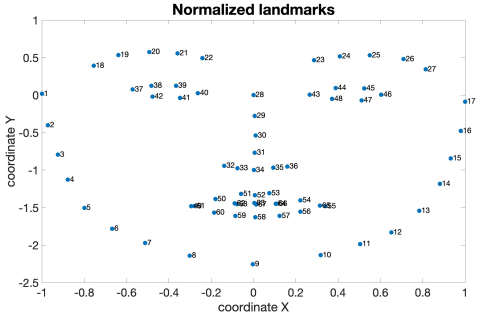
The position of these landmarks is highlighted with four red circles in Figure 3.14a, wherein the coordinates are reported in the plane prior to their normalization.

So, the distances between pairs of these points were used to normalize the coordinates of all 68 landmarks. Specifically, the horizontal distance between the left meningeal point and the nasal centre point was used to normalize the x coordinates of all points to the left of the nasal centre point. Similarly, the x coordinates of all points to the right of the nasal centre point were normalized using the horizontal distance between this point and the right meningeal point. The use of two separate distances for the horizontal normalization of the points on the left and right of the nasal centre point allowed to balance the symmetry between the right and left portions of the face. The vertical distance between the nasal centre point and the subnasal centre point was used to normalize the y coordinates of all points, regardless of their position.

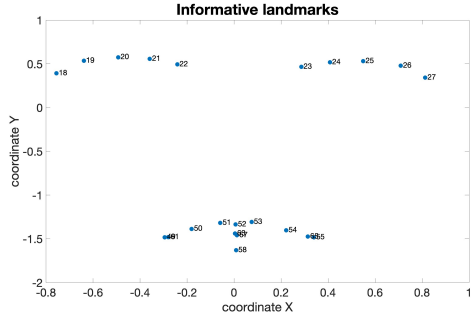
After applying coordinate normalization, a selection procedure was implemented to identify a reduced number of points deemed to be highly informative. Among all existing AUs, the FACS manual [155] isolates 22 considered fundamental as they describe specific facial muscles responsible for human facial expressions. In particular, the importance of these 22 facial muscles for understanding human expressions is highlighted as they are responsible for the contraction and relaxation of different parts of the face, thus their movement is closely related to facial expressions. For these reasons, only the coordinates of the 22 points related to these specific muscles, and their movement, were considered for the realisation of the database. This allowed for the isolation of only the most significant points for the considered applications, resulting in a highly performant database optimised for its specific purpose. The aforementioned steps can be observed in the sequence of images presented in Figure 3.14.



(a) Raw data.



(b) Normalized data.



(c) Informative data.

Figure 3.14. Facial landmarks coordinates elaboration process: (**a**) coordinates of the landmarks as obtained by the algorithm under investigation; (**b**) coordinates after applying normalization; (**c**) coordinates of the 22 key points selected for their high degree of informativeness.

The described procedure was implemented following a meticulous performance evaluation, which involved employing classification algorithms to

assess the dataset's efficiency, as detailed in the following section. The evaluation conducted revealed the presence of samples with limited informative content regarding emotional expressiveness. More precisely, including such samples introduced significant bias, compromising performance. Consequently, we decided to reduce the number of features associated with each sample, considering only 22 landmarks. This reduction not only improved performance but also alleviated computational load, enabling faster analysis.

This database, named *FeelPix*, is available on GitHub.com [203] for other researchers and developers to use for the development of facial expression recognition algorithms based on landmarks.

Compared to existing datasets for facial expression recognition, which primarily consist of images, our dataset provides detailed data based on facial landmarks, eliminating the need for further processing through Convolutional Neural Networks or facial landmark coordinate extraction. Moreover, while most existing datasets have labels generated by developers, FeelPix stands out for collecting the data labels directly from the participants involved in the study, minimizing the potential errors or biases introduced by third-party labeling that did not personally experience the captured emotional states during detection.

Testing Algorithm To assess the effectiveness of the developed dataset, a simple Facial Expression Recognition algorithm was created and trained using this data. The dataset was processed to reduce the classification problem to a binary one, where:

- The selection of a specific emotion was associated with class 1;
- The absence of the specific emotion's selection was associated with class 0.

In this manner, an algorithm employing seven machine learning classifiers, each optimised for a specific emotion, was developed. The classifiers utilized in this study include Support Vector Machine (SVM) and Random

Forest (RF), chosen for their efficiency and lower data requirements compared to neural networks.

However, it emerged that the number of samples between the two classes was highly imbalanced, with a larger quantity of samples in class 0 compared to class 1. This imbalance is attributed to class 1 representing the selection of the specific considered emotion, while class 0 encompasses the non-selection of that specific emotion but the possibility of selecting all other emotions. Consequently, before training the classifiers, a final dataset processing step was performed to mitigate the extreme imbalance. The undersampling approach was adopted, keeping all data in the minority class and reducing the size of the majority class.

This approach was employed by considering the number of positive samples in the dataset under examination and randomly selecting a specific number of negative samples, thus facilitating the construction of a less imbalanced dataset. In particular, since the negative samples represent a greater variety of emotions than the positive samples, a partial approach was adopted: the difference in samples between the two classes was appropriately regulated and set to half the number of samples present in the minority class.

Subsequently, an optimisation of both considered classification methodologies was carried out for each emotion, through a random search of hyperparameters. Lastly, both methodologies were trained on the developed database, to determine which of the two ones yielded the best performance for each emotion.

To achieve this, a five-fold cross-validation was conducted, considering accuracy, precision, F-measure, and G-mean as validation metrics. Precision, F-measure, and G-mean were chosen because they are insensitive to the dataset's imbalance, which, although mitigated through undersampling, still exhibited slight imbalances.

Results

In our study, we had three primary objectives. Firstly, we aimed to develop a universal methodology for evaluating the performance of Facial Expression Recognition algorithms. Subsequently, we collected labeled data during tests to create a comprehensive dataset. Finally, we focused on developing an FER algorithm capable of verifying the reliability of the created database by classifying facial expressions based on facial landmark coordinates.

The results will demonstrate the achievement of these objectives by showcasing the algorithm's performance using the developed validation protocol. Additionally, the validity of the released dataset and the proposed algorithm will be highlighted by presenting the algorithm's performance on the dataset's data. Furthermore, a comparison between the two performances will be provided.

Investigated Algorithm Results To assess the reliability of the investigated FER algorithm, i.e., the one implemented in the face-api.js library, the success rate for each user was initially calculated by comparing their selected emotions with the library's detections for each image. Specifically, this quantity was determined as the ratio between the number of images in which at least one of the user's choices and at least one of the algorithm's detections agreed with each other, and the total number of images displayed to the subject (i.e., 35 images).

Unexpectedly, the results obtained were highly variable among different users and, on average, below 54%. Given the significant deviation from the expected outcome, it was deemed appropriate to calculate the same validation metrics used to evaluate the testing algorithm's performance for each emotion.

However, as reported in Table 3.5, even these values were considerably lower than expected: the metrics assume percentages below 40% in most cases and never exceed 60%, as shown in Figure 3.15.

Table 3.5. Validation metrics for the algorithm under examination.

	Anger	**Disgust**	**Fear**	**Happy**	Neutral	**Sad**	**Surprise**
Precision	24%	41%	14%	25%	35%	32%	55%
F score	11%	9%	3%	35%	52%	35%	33%
G mean	26%	22%	12%	60%	4%	51%	47%

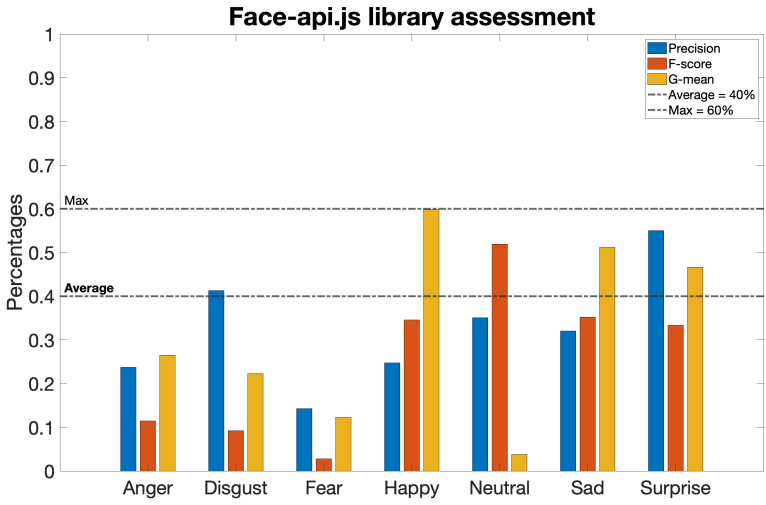


Figure 3.15. Validation metrics—Precision, F-score, and G-mean—pertaining to the investigated algorithm during the validation process. These metrics are presented for each discrete emotion category, including neutrality.

Outcomes of the Dataset Testing The FeelPix dataset was utilized to train and test the algorithm developed for its evaluation. This enabled determination of the performance achievable using such data. In order to conduct a comprehensive investigation, validation metrics were computed for both considered classification methodologies (i.e., SVM and RF).

This facilitated identification of the methodology exhibiting the best performance for each emotion, thereby composing the algorithm that recog-

nises all seven emotions from the database data. The validation outcomes of the various classifiers on the created database dataset met the expectations, as the metrics exhibit values higher than 75% in most cases and never lower than 64%, as evident from Table 3.6 and Figure 3.16, where the values associated with the best performing classifier for each emotion are presented.

Table 3.6. Validation metrics for the testing algorithm.

	Anger	**Disgust**	**Fear**	**Happy**	Neutral	**Sad**	**Surprise**
Accuracy	78%	78%	72%	65%	67%	74%	80%
Precision	81%	85%	74%	65%	75%	79%	85%
F score	81%	83%	77%	65%	72%	78%	82%
G mean	77%	74%	71%	64%	65%	74%	80%

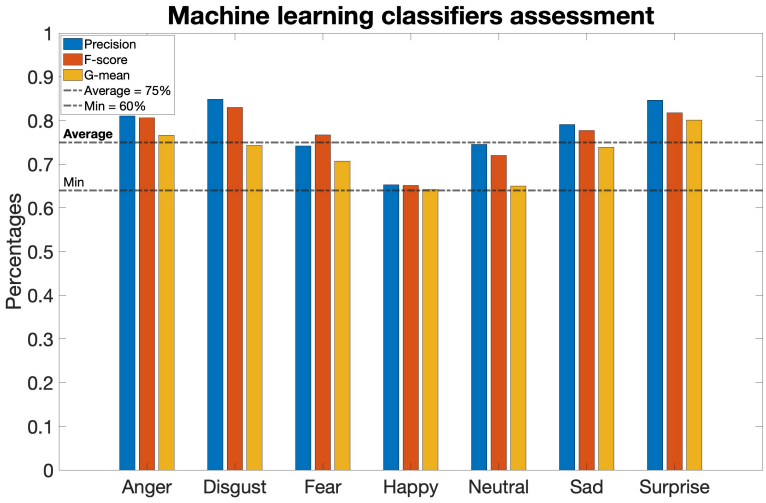


Figure 3.16. Validation metrics acquired from the algorithm developed to ascertain the integrity of the proposed FeelPix database. These metrics include precision, F-score, and G-mean, and are displayed for each emotion category.

FER Algorithms' Comparison It is evident that the validation metric values obtained using the combination of the developed database and proposed testing algorithm are significantly better than the results shown by the FER algorithm under investigation, as depicted in the Figure 3.17.

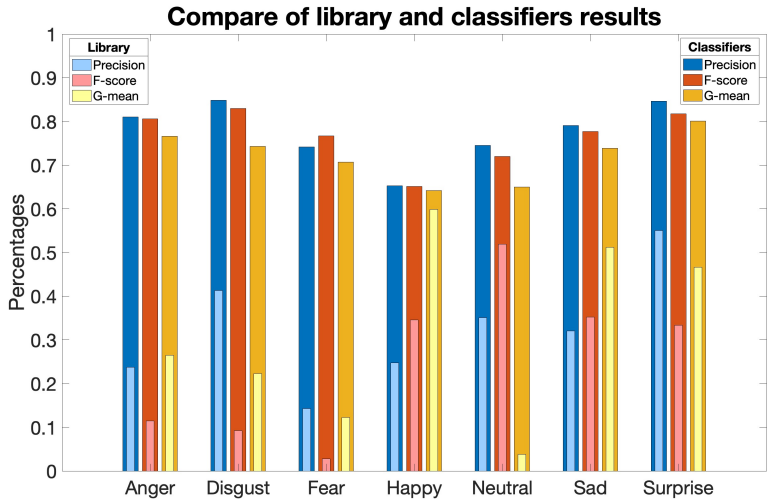


Figure 3.17. Comparing the outcomes produced by the algorithm under investigation during the validation process with those achieved by the algorithm developed for validating the proposed database. The visualization includes precision, F-score, and G-mean metrics, shedding light on the performance of each algorithm across diverse emotion categories.

Conclusively, the developed and released database proves to be efficient and functional, as it achieves sufficiently high performance even with a simple classification algorithm.

Discussion

In this work, we have investigated the world of emotions, specifically their recognition through facial expression analysis. In light of the current state-of-the-art in emotion recognition, it has emerged that detecting emotions in naturalistic conditions still presents significant difficulties. Numerous variables contribute to these difficulties, such as the lack of a uni-

versally accepted definition of facial codes and/or facial actions. Furthermore, the comprehensive understanding and prediction of affective processes undoubtedly requires careful integration of multiple contextual factors, information modalities, and evaluations within naturalistic environments. Therefore, evaluating and verifying FER algorithms is a challenging but critical task.

For this reason, we developed a protocol to validate FER algorithms in this work. In the protocol, participants were shown static emotional stimuli on a computer screen and asked to select the emotion they felt while the facial expression was detected by the FER algorithm. The test was performed through a web application, which proved to be functional, easy to use, intuitive, and visually appealing. Furthermore, the application did not encounter any problems during data acquisition, and the results were correctly saved for all subjects in the selected database. Therefore, the developed system can be considered an effective tool for validating FER libraries. Furthermore, the proposed technology also enables the creation of a database of spontaneous expressions that incorporates various contextual factors, such as different genders, ethnicities, personalities, and cultures.

The validation of the FER algorithm implemented in the JavaScript library face-api.js used in this study yielded results that were not in line with the initial expectations. In fact, the algorithm's performance was significantly inferior to the generally accepted values, indicating a poor ability to correctly recognise facial expressions. However, it is important to consider that these results are limited by the context of the type of stimulus and the experimental setup. Additionally, it is worth noting that the results obtained may also be influenced by the selection of participants and their degree of expressiveness.

On the other hand, the labelled data database, created by processing the landmark coordinates provided by the algorithm under examination and the user choices, proved to be accurate, functional, and reliable, as demonstrated by the validation metrics achieved by the testing algorithm applied on it. In fact, the combination of the proposed testing algorithm and the

developed database allows for the recognition of all seven emotions with high accuracy, making it a valuable tool in the field of behavioural sciences and facial expression recognition.

It is important to note that the data used to create the database were collected under experimental conditions where a constant level of expressiveness of the subjects was not guaranteed, thereby generating a dataset that covers a wide range of conditions. Therefore, considering that this database has successfully captured even the most subtle facial expressions, it can be deemed as a valuable asset for a range of potential applications, such as human-computer interaction, affective computing, and mental health diagnosis. Furthermore, in future developments, minor changes, such as expanding the type of stimulus or the experimental environment, would allow more information to be added to the database, thus enabling the proposal of a tool that can integrate the multiple contextual factors that typically make facial expression recognition challenging.

Conclusions

In conclusion, we have presented a comprehensive validation system for Facial Expression Recognition (FER) algorithms, offering several significant advantages. Firstly, the system's applicability to any type of FER algorithm allows for the efficient determination of its effectiveness, reducing the algorithm verification phase, and ensuring the utmost accuracy and reliability in identifying and interpreting facial expressions. Moreover, the system provides a clear and transparent description of the algorithm's proficiency in recognising and interpreting facial expressions, facilitating the identification of any potential bias or errors. Additionally, the validation system supports the development of new solutions and applications, fostering a deeper understanding of its capabilities and limitations, which is invaluable for advancing the field of FER.

Furthermore, as part of our contributions, we have developed and released a meticulously labelled data database, which bestows various advantages. The database offers extensive coverage of a wide range of conditions,

encompassing both highly expressive and less expressive subjects, providing a rich and diverse dataset for research and experimentation. Accessible and user-friendly, the database empowers researchers and professionals to readily employ it in their studies and application development pertaining to emotion recognition through facial expression analysis.

Finally, the algorithm we have devised for database verification enables the precise identification and interpretation of emotions associated with facial expressions, offering profound insights into individuals' emotional reactions. Consequently, it proves instrumental in facilitating more natural and intuitive interactions between humans and technological interfaces, such as robots, virtual assistants, and augmented reality systems, making it a powerful tool for psychological and social research.

By combining these contributions, our work advances the frontiers of facial expression recognition, paving the way for enhanced emotional understanding and human-computer interactions in various domains.

3.6 Enhancing Emotional Congruence in Sensory Substitution

Introduction

Sensory Substitution Devices Sensory substitution devices (SSDs) represent a non-invasive approach to compensating for sensory loss by translating information from one sensory modality to another [204]. These systems operate on the principle that the brain processes sensory information as electrical signals, independent of their source, enabling the interpretation of sensory data through alternative channels [205]. This neurological adaptability, termed neuroplasticity, allows areas of the brain typically associated with one sense to process information from different sensory inputs [206].

The fundamental architecture of SSDs comprises three primary components: sensors that capture environmental information, a coupling device that processes and converts these signals, and actuators that stimulate the substituting sensory modality [204]. This framework supports various sensory translations, including vision-to-tactile, audio-to-tactile, and vision-to-audio conversions.

A seminal development in this field was Paul Bach-y-Rita's Tactile Vision Substitution System (TVSS) in 1969 [207]. The TVSS utilised a 20x20 array of vibrotactile stimulators to convert visual information into tactile patterns on the user's back. This pioneering work demonstrated that subjects could learn to interpret complex visual information through tactile stimulation, particularly when actively controlling the input device [208].

Contemporary devices have evolved significantly in both sophistication and portability. The BrainPort, developed by Bach-y-Rita in the 1990s [209], exemplifies this progression through its use of electrotactile stimulation on the tongue, chosen for its high receptor density and conductive properties. The device employs a direct spatial mapping strategy, where pixel brightness correlates with stimulation intensity, enabling users to develop capabil-

ities such as object recognition and spatial navigation following appropriate training [210].

In the audio-visual domain, the vOICe system [211] pioneered the conversion of visual information into auditory signals through a systematic encoding algorithm. The system maps vertical position to frequency, horizontal position to temporal sequence, and brightness to sound amplitude. More recent developments, such as the EyeMusic [212], have enhanced this approach by incorporating colour information through varying musical instruments, demonstrating the potential for more nuanced sensory translation.

Significant advances have also emerged in tactile-hearing substitution, particularly for individuals with hearing impairments. Modern devices like the Neosensory Buzz [213] utilise targeted vibrotactile feedback to assist in phoneme recognition, specifically processing high-frequency speech components that are often problematic for individuals with hearing loss. The system employs four actuators, each corresponding to specific phonemes (/s/, /t/, /z/, and /k/), activating for 80 ms when the respective sound is detected.

The implementation of multimodal feedback represents a particularly promising direction in SSD development. The Sound of Vision (SoV) system [214] combines auditory and vibrotactile feedback to provide comprehensive environmental information, capable of detecting obstacles up to 3.5 metres away. This multimodal approach aligns with the natural integration of sensory information in human perception [6].

Building on these advances in sensory substitution technology, our work addresses a critical gap in audio-visual SSDs: the limited transfer of emotional content from visual scenes. While current devices effectively translate spatial and temporal information, they struggle to convey the emotional essence of visual experiences. We propose a novel approach that combines semantic video segmentation with user emotional response analysis to develop musical features for video sonification. Through systematic experimentation and data analysis, we demonstrate how emotion-aware musical

mappings can enhance the affective dimension of sensory substitution.

Current Challenges and Limitations A significant challenge in sensory substitution technology centres on the limited transfer of emotional content, an issue first identified by Bach-y-Rita et al. [14]. While recent advances have improved functional information transfer, emotional engagement remains problematic. Modern SSDs like Eagleman's haptic devices [205] and Goral's interoceptive-exteroceptive systems [215] demonstrate progress in emotional transfer, yet face persistent challenges.

Current limitations manifest in several ways: visual-emotional translation remains incomplete, particularly in facial expression recognition and artistic appreciation [205]; social interaction barriers persist due to limited emotional feedback [215]; and prosthetic interfaces, despite biological integration advances [216], struggle with affective dimension transfer.

Recent approaches to address these limitations include haptic emotional recognition systems for ASD users and interoceptive-exteroceptive substitution for enhanced emotional regulation [215]. However, these solutions still fall short of natural emotional perception, particularly in social contexts where emotional nuance is crucial.

The emotional deficit in sensory substitution manifests in several ways. While devices can accurately translate spatial and temporal information, they often fail to convey the subtle nuances that contribute to emotional resonance in natural sensory experiences. For instance, visual-to-auditory devices may successfully communicate object location and form but struggle to convey the emotional impact of facial expressions or artistic works.

Beyond the emotional limitations, several practical challenges impede widespread SSD adoption. The significant cognitive and emotional investment required during training periods presents a substantial barrier to user acceptance [6]. Additionally, the financial implications of both device acquisition and necessary professional training support create accessibility challenges [6].

Current research directions emphasise developing more accessible and

flexible devices that can be customised to individual user needs, potentially increasing autonomous usage and acceptance rates. Kristjánsson [6] proposes several critical design considerations: maintaining unobstructed access to remaining natural senses, ensuring minimal interference with user mobility, reducing training requirements, and carefully managing information density to prevent sensory overload.

Related Work

Music and Emotional Response Research on music-emotion relationships has established robust connections between musical features and emotional responses. Studies demonstrate consistent correlations between structural elements like tempo, mode, and dynamics with specific emotional states [217]. Physiological measurements validate these relationships, showing reliable patterns in heart rate, skin conductance, and facial muscle activity in response to emotional music [218].

Recent work has expanded our understanding of musical emotion processing through neuroimaging studies. [219] identified distinct neural networks involved in processing different musical emotions, while [220] established the universality of basic musical emotion recognition across cultures. These findings suggest a biological basis for music-emotion associations, though cultural factors modulate specific responses [221].

Visual Features and Emotional Processing Visual emotion processing research reveals systematic relationships between visual properties and emotional responses. Colour characteristics strongly influence emotional perception, with brightness and saturation particularly affecting valence and arousal dimensions [222]. These associations demonstrate remarkable consistency across cultures while maintaining some cultural specificity [223].

Contemporary studies have advanced our understanding of visual emotion processing through computational approaches. [224] established correlations between geometric properties and emotional responses, while [225] developed algorithms for predicting emotional responses to images based

on low-level visual features. Facial expression recognition research has particularly benefited from deep learning approaches, achieving human-level performance in emotion classification [202].

Cross-modal Emotional Integration Cross-modal emotional processing involves complex interactions between sensory modalities. Neuroimaging studies have identified key brain regions, particularly the superior temporal sulcus and anterior cingulate cortex, in integrating emotional information across modalities [181]. Behavioural research demonstrates that emotional congruence between modalities enhances perception and recognition [11].

Recent studies have explored temporal aspects of cross-modal integration. [226] found that emotional congruence affects early sensory processing, while [227] demonstrated that temporal synchrony between auditory and visual emotional signals enhances integration effectiveness. These findings have important implications for SSD design, suggesting the need for precise temporal alignment in emotional content translation.

Emotional Processing in Current SSDs While existing SSDs successfully translate basic sensory information, emotional content preservation remains challenging. Traditional systems like vOICe and EyeMusic focus primarily on spatial and temporal information transfer [211, 212]. Recent attempts to incorporate emotional mapping have shown promise but remain limited in scope and effectiveness.

Contemporary approaches have begun addressing this limitation. [205] developed haptic systems incorporating emotional feedback, while [215] introduced interoceptive-exteroceptive substitution for enhanced emotional engagement. However, systematic evaluation of emotional transfer effectiveness remains scarce, particularly regarding long-term user engagement and acceptance.

Research Questions

While prior work has established relationships between musical features and static visual stimuli like colours [223], this study investigates emotional coherence in audiovisual sensory substitution using dynamic visual content. Building on existing SSD limitations in emotional transfer [14], we present a novel experimental protocol combining validated emotional music samples with video stimuli, while recording physiological responses and facial expressions to verify emotional coherence. This systematic approach aims to develop emotion-aware mappings for more engaging sensory substitution systems.

The first research question explores how visual stimuli of varying emotional content map to musical selections in the valence-arousal space:

RQ1: To what extent do valence and arousal dimensions of emotional responses drive the association between visual stimuli and musical features?

This investigation will examine whether participants predominantly match visual stimuli to music based on valence alignment, arousal correspondence, or a combination of both dimensions, as measured through the SAM scale.

The second research questions aims to verify the emotional consistency between the physiological and facial expression recordings and the SAM:

RQ2: Can psychophysiological responses accurately predict SAM ratings?

This research question aims to validate the experimental protocol by examining the alignment between participants' subjective self-assessments and their objective psychophysiological responses. Strong predictive relationships would confirm the internal consistency of our emotional measurements

The third research question addresses the predictive aspects of these emotional associations:

RQ3: Can a machine learning system effectively predict appropriate musical features for visual stimuli based on emotional response data?

This explores the feasibility of developing an automated system that selects musical accompaniment for visual content by considering both the

objective features of the visual stimulus and its emotional impact, as quantified through systematic emotional response measurements.

These questions aim to establish a foundation for emotion-aware sensory substitution systems that can maintain emotional coherence in the translation from visual to auditory modalities. The findings will contribute to the development of more engaging and emotionally appropriate sensory substitution solutions.

Study Overview

This study presents a novel approach to investigating emotional congruence in sensory substitution systems through four main contributions. First, we introduce a new experimental protocol for measuring emotional associations between visual stimuli and musical characteristics. This protocol systematically evaluates participants' emotional responses using the Self-Assessment Manikin (SAM) scale [136] while they experience various combinations of visual stimuli and musical accompaniment. Second, we present a comprehensive database collected from 36 participants, containing emotional response measurements across multiple visual-auditory pairings. The database includes both SAM ratings and detailed annotations of the musical characteristics associated with different emotional responses. Third, we provide a statistical analysis of the collected data, focusing on the relationship between visual content, musical features, and emotional responses in the valence-arousal space. Finally, we develop and evaluate a predictive model that leverages this emotional response data to automatically select appropriate musical characteristics for given visual inputs, considering both content features and emotional impact.

Experimental Protocol

This study aimed to quantify emotional coherence between visual and auditory stimuli through behavioural responses and physiological measurements. The protocol was designed to systematically evaluate participants'

emotional responses to video clips under varying audio conditions, whilst recording facial expressions and autonomic nervous system activity. The experimental design enabled direct comparison between subjective ratings and objective physiological markers of emotional response.

This study underwent ethical evaluation by King's College London University's ethical committee and received approval in November 2023.

Participants Thirty-six healthy participants (26 female [72.2%], 10 male [27.8%]) were recruited for this study, with a mean age of 25.94 years (SD = 8.95, range 20-57). Ten participants (27.8%) reported wearing corrective lenses. All participants met the following inclusion criteria: normal or corrected-to-normal vision, no hearing impairments, no history of neurological conditions, and no previous exposure to the selected film clips. The study was conducted in accordance with the Declaration of Helsinki, and all participants provided written informed consent before participation. One participant was excluded from the analysis due to incomplete data collection resulting from non-adherence to the experimental procedure.

Experimental Design The experiment followed a repeated measures design, with each participant viewing four video clips across multiple audio conditions while physiological signals (heart rate, electrodermal activity) and facial expressions were continuously recorded. Sessions lasted 30-45 minutes and were conducted individually in a controlled laboratory environment.

Stimuli Four video clips, each 30 seconds in duration, were selected as visual stimuli. For each video, participants experienced four distinct conditions: one silent viewing and three audio accompaniments. The audio conditions were systematically rotated among the following categories:

- Emotionally congruent audio (matching both valence and arousal)
- Partially congruent audio (matching either valence or arousal)

- Emotionally incongruent audio

Procedure Each experimental session followed a structured protocol:

- Initial Setup: upon arrival, participants were briefed about the experimental procedure and provided informed consent. The experimenter then fitted participants with the Empatica E4 device for physiological measurements and Sennheiser HD 450BT over-ear headphones for audio playback. The laptop's integrated webcam was positioned to record participants' facial expressions throughout the session. A 3-minute baseline physiological recording was conducted while participants were at rest.
- Stimulus Presentation: each trial began with the silent viewing of a video clip, followed by three presentations of the same clip paired with different audio conditions. Between each stimulus presentation, a 30-second rest period was implemented to allow physiological measures to return to baseline.
- Response Collection: following each stimulus presentation (both silent and with audio), participants completed the Self-Assessment Manikin (SAM) scale to rate their emotional state in terms of valence and arousal. Prior to the experiment, participants received comprehensive instructions on using the SAM scale. After experiencing all conditions for each video, participants indicated which image-sound pairing they found most effective. The interfaces shown to subjects can be seen in Figure 3.19.
- Inter-stimulus Protocol: to ensure reliable data collection and participant comfort:
    1. Physiological signals were continuously monitored.
    2. A 30-second rest period was enforced between stimuli.
    3. A 30-second rest period was enforced after the 4 showings of each video.

All sessions were conducted in a controlled laboratory environment with standardised lighting and audio presentation conditions.

Figure 3.18. Experimental setup showing participant using the tablet for SAM scale ratings while wearing Empatica E4 wristband for physiological measurements and Sennheiser HD 450BT headphones for audio playback. The laptop screen displays video stimuli while the integrated webcam records facial expressions.

Stimulus Selection and Characteristics

Visual Stimuli Four visual stimuli were selected from the DECAF database [228], a resource that offers validated emotional ratings and corresponding physiological responses for a range of movie clips. The selection process focused on identifying clips that most strongly represented the four extremes of the valence-arousal space: Low Valence-Low Arousal (LVLA),

Low Valence-High Arousal (LVHA), High Valence-High Arousal (HVHA), and High Valence-Low Arousal (HVLA). The final selection was based on the clips showing the most polarized ratings within each quadrant (Table 3.7).

Selected clips span a representative range of the valence-arousal space, with four distinct emotional quadrants represented (Table 3.7). Each 30-second clip was chosen to elicit a specific primary emotion while maintaining natural viewing conditions through the use of commercially released films.

Category	Valence		Arousal		Scene Description
	μ	σ	μ	σ	
LVLA	-0.85	0.62	-0.82	1.06	Young girl cries at friend's funeral
LVHA	-1.24	0.73	1.20	0.88	Lady accidentally dies during magic act
HVHA	0.99	0.63	1.15	0.88	Passengers react to pilot's struggle
HVLA	0.76	0.68	-1.12	1.02	Son meets birth mother at concert

Table 3.7. Characteristics of Selected Video Stimuli.

Audio Stimuli Audio stimuli were selected from the International Affective Digitized Sounds (IADS) [135] database, which provides standardised affective ratings for naturally occurring sounds. Seven audio clips were chosen to represent varied points in the valence-arousal space (Table 3.8), enabling different levels of emotional congruence with the visual stimuli.

The range of selected audio stimuli (arousal: 2.23-7.92; valence: 2.56-7.75) ensures coverage of multiple emotional quadrants, allowing for both congruent and incongruent audiovisual pairings. All ratings are on a scale from 1 to 9, where higher values indicate greater arousal and more positive valence respectively.

Sound ID	Arousal [1, 9]	Valence [1, 9]
0010	4.83	5.63
1100	7.91	6.59
1114	7.92	7.75
1122	2.23	7.50
1128	2.75	6.33
1157	2.95	5.86
1392	7.59	3.14

Table 3.8. Characteristics of Selected Audio Stimuli.

Data Acquisition and Processing

Data Streams Three synchronized data streams were collected during the experimental sessions (Table 3.9): physiological measurements from the Empatica E4 device, facial expressions captured via webcam, and visual content features extracted from the stimulus videos. Each stream underwent specific processing methods to extract relevant features for the emotion recognition system. The detailed processing methodology for each data stream is described in the following subsections.

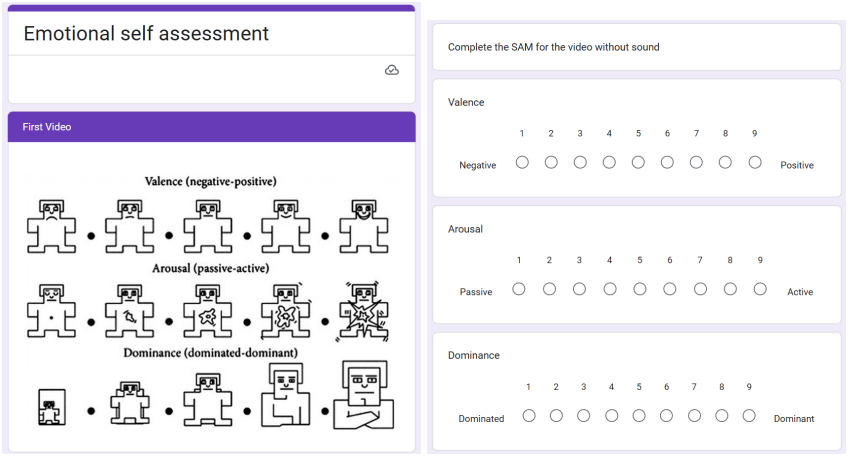
Data Stream	Features	Processing Method
Physiological	Heart Rate (HR)	Peak detection after baseline removal
	Heart Rate Variability (HRV)	RMSSD from inter-beat intervals
	Galvanic Skin Response (GSR)	CVX optimisation, artifact removal
Facial	Facial Action Units	PyFeat real-time FAU detection
	Facial Features	Temporal alignment with content windows
Visual	Frame Features	MineCLIP feature extraction
	Content Windows	L1 norm temporal segmentation

Table 3.9. Data Stream Processing Overview.

While peripheral temperature was initially recorded, preliminary analysis revealed that temperature variations during the experimental sessions

were too small to provide meaningful information for emotion recognition. Therefore, temperature data was excluded from subsequent analyses.

(a) Baseline screen with countdown timer displayed during the 3-minute physiological recording period, allowing synchronization between participant responses and wearable data collection.



(b) SAM interface presented after each video, displaying pictograms for three emotional dimensions (valence, arousal, dominance) with 9-point rating scales.

Figure 3.19. Interface screens showing baseline recording phase (a) and post-stimulus SAM assessment (b). Additional screens follow similar format with countdown timers to synchronize physiological recordings

Signal Processing Pipeline The processing pipeline integrates multiple stages of data analysis and feature extraction (Figure 3.20). The visual processing begins with MineCLIP application to extract frame-level features

from the input video [229]. These features serve to identify significant content changes in the visual stream, utilizing L1 norm calculations for temporal segmentation. This segmentation process establishes the fundamental windows that guide the subsequent analysis of all data streams.

The physiological signal processing encompasses two main components. For the PPG signal, we first remove baseline drift through CVX optimisation, followed by peak detection and validation. These processed peaks enable the calculation of heart rate from inter-beat intervals and the computation of heart rate variability (RMSSD). The GSR signal undergoes similar initial processing, with baseline correction through CVX optimisation, followed by artifact removal and response detection and quantification.

All physiological signals are then analysed within the windows identified from the visual content analysis. For each window, we compute statistical features including maximum value, minimum value, and standard deviation, providing a comprehensive representation of the physiological response patterns during each segment of visual content.

Facial expression analysis runs parallel to these processes, with PyFeat performing continuous detection of facial action units throughout the video recording. These expressions are temporally aligned with the previously identified content windows, and statistical features are extracted for each window, maintaining consistency with the physiological feature extraction approach.

Musical Feature Extraction Musical features were extracted using MIRToolbox in MATLAB R2020a, focusing on five key characteristics (Table 3.13). These features were selected based on their established relationship with emotional perception in music, as discussed previously. Zero crossing rate and RMS provide temporal information about signal intensity and roughness, while spectral centroid captures timbral brightness. Mode and key detection offer harmonic information, which is crucial for emotional valence assessment. Frame-level processing used two different window sizes based on feature requirements. Temporal features (zero crossing, RMS)

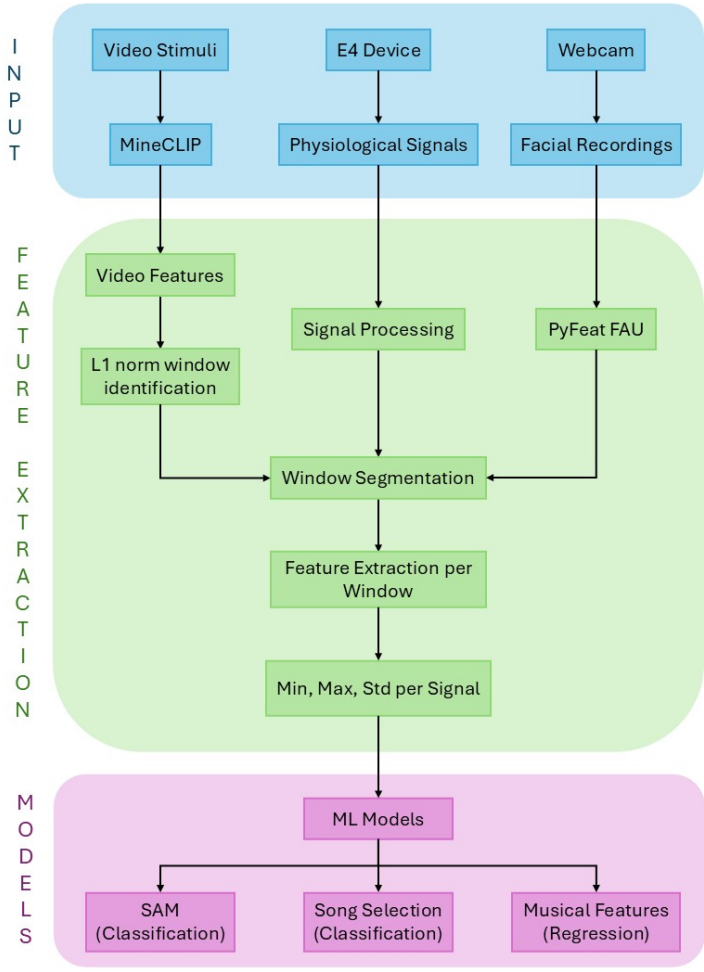


Figure 3.20. Signal processing pipeline. The system processes three parallel input streams: video content through MineCLIP feature extraction, physiological signals from the E4 device (PPG, GSR), and facial expressions via webcam recording. MineCLIP features are used to identify temporal windows through L1 norm analysis, which are then applied to segment the physiological and facial expression data. For each window, statistical features are extracted from all data streams for subsequent machine learning analysis.

were computed using 50ms frames with 50% overlap to capture rapid signal changes. Harmonic features (mode, key) required longer 1-second frames with 0.5-second hop factor to ensure sufficient information for reliable tonal analysis.

Feature	Description
Zero cross*	Counts the number of times the signal crossed the X-axis, indicating noisiness
Centroid	Returns the centroid of the data, describing the spectral distribution
Mode**	Estimates the modality (major vs. minor) on a scale from -1 to +1, where +1 indicates major and -1 indicates minor
Key**	Provides broad estimation of tonal centre positions by returning the best key(s)
RMS*	Computes the global energy of the signal by taking the root average of the square of the amplitude

* Frame length: 50 ms, 50% overlap
** Frame length: 1 s, 50% overlap (0.5 s hop factor)

Table 3.10. Musical Features Extracted for Analysis.

Analysis Methodology

Emotional Mapping Distribution To analyse how participants matched visual stimuli to musical selections, we computed the percentage distribution of choices based on valence and arousal alignment using SAM ratings. This analysis directly addresses RQ1 by quantifying whether selections were driven primarily by valence, arousal, or both dimensions.

Emotional Response Validation To validate the coherence between self-reported and measured emotional responses, we implemented classification

models predicting SAM ratings (high/low values) from physiological and facial data. Each participant completed the SAM assessment for all four 30-second video clips, each presented with four different sound conditions, resulting in 16 SAM scores per participant. With 35 viable participants, this yielded a total of 560 samples for analysis. By evaluating different combinations of input features (physiological signals, facial expressions, and video content features), we aimed to understand which data sources best capture emotional states and whether combining multiple modalities improves prediction accuracy. We compared Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), and K-nearest neighbours (kNN classifiers), using accuracy as the performance metric. The following feature combinations were evaluated:

- Physiological + Facial Expression features
- Physiological + Video features
- Facial Expression + Video features
- All features combined

This systematic comparison helps identify which emotional response channels provide complementary information and which combinations yield the most reliable predictions of self-reported emotional states.

Musical Feature Prediction We first attempted to directly predict song selection using the same multimodal feature combinations and classification algorithms used for emotional response validation. However, the poor performance of this approach suggested that specific musical characteristics, rather than complete songs, drive user preferences. With 35 participants each selecting one song for each of the four 30-second video clips, we obtained a total of 140 samples for this analysis. Therefore, to address RQ2's goal of automated music selection for sensory substitution devices, we developed regression models to predict individual musical features (zero

crossing, centroid, mode, key, RMS) from different combinations of emotional response data:

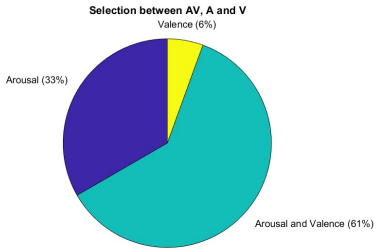
- Physiological + Facial Expression features
- Physiological + Video features
- Facial Expression + Video features
- All features combined

This systematic evaluation aimed to identify which input modalities are most crucial for predicting appropriate musical features, informing the minimum sensor requirements for effective SSD design. We compared Random Forest Regressor (RFR), XGB Regressor (XGBR), Support Vector Regressor (SVR), and Linear Regression (LR), using MSE for evaluation.

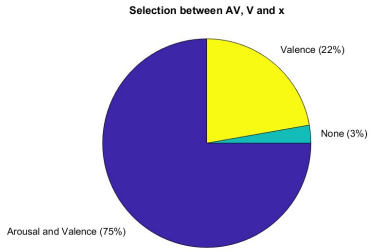
Model Training and Evaluation Protocol All models followed a nested 5-fold cross-validation procedure. The outer loop performed a 5-fold cross-validation, using one fold for testing and the remaining four for training. Inside each training split, a 5-fold cross-validation was applied to optimise hyperparameters using GridSearchCV. Features were standardized using StandardScaler, applied only to the training data within each fold to avoid data leakage. Hyperparameter optimisation was performed through grid search with cross-validation on the training set, with final evaluation on the held-out test set of each outer fold. Initial dimensionality reduction experiments with PCA and t-SNE showed that raw features provided optimal performance, so the final models used the complete standardized feature set. For each model, we report both the mean performance and standard deviation across the outer cross-validation folds to demonstrate the robustness of our results.

Experimental Results

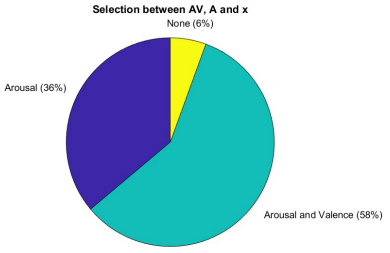
Emotional Mapping Distribution Participants' preferences for audio-visual emotional coherence showed clear patterns across stimulus combinations



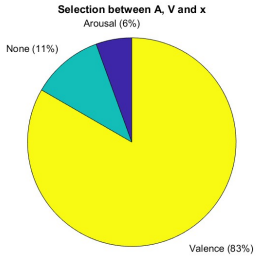
(a) Selection distribution between coherent (AV), arousal-matched (A), and valence-matched (V) stimuli



(b) Selection distribution between coherent (AV), valence-matched (V), and incoherent (X) stimuli



(c) Selection distribution between coherent (AV), arousal-matched (A), and incoherent (X) stimuli



(d) Selection distribution between arousal-matched (A), valence-matched (V), and incoherent (X) stimuli

Figure 3.21. Distribution of participants' preferences across different combinations of emotionally coherent and incoherent audio-visual pairings. AV indicates stimuli matching in both arousal and valence, A indicates arousal-only matching, V indicates valence-only matching, and X indicates no emotional coherence.

(Figure 3.21). For fully coherent stimuli (matching both arousal and valence, AV), participants strongly preferred maintaining alignment in both emotional dimensions (61-75% of selections), compared to matching only arousal (33-36%) or only valence (6-22%).

When presented with partially coherent options, a significant asymmetry emerged between valence and arousal matching. In conditions contrasting arousal-matched stimuli (A) with valence-matched ones (V), partici-

pants overwhelmingly selected valence-matched options (83%), with only 6% choosing arousal-matched alternatives (Table 3.21). This finding suggests valence coherence plays a more crucial role in perceived audio-visual emotional alignment than arousal coherence.

Emotional Response Validation Initial analysis of SAM ratings revealed distinct patterns across emotional dimensions. Dominance showed a strong preference for high ratings (65.2%), while valence responses were more balanced between low (54.8%) and high (45.2%) values. Arousal ratings demonstrated a slight tendency toward high activation states (54.5%).

Valence				
Feature Set	**RF**	**XGB**	**KNN**	**SVM**
Physiological	53.9 ± 3.8	51.6 ± 4.5	51.8 ± 5.4	54.8 ± 3.8
Facial expressions	62.0 ± 4.7	62.0 ± 4.0	61.6 ± 4.8	63.0 ± 3.6
Physiological + Facial expressions	61.4 ± 6.3	58.8 ± 6.6	51.6 ± 5.5	59.8 ± 5.2
Arousal				
Feature Set	**RF**	**XGB**	**KNN**	**SVM**
Physiological	57.3 ± 4.6	58.0 ± 4.9	46.1 ± 4.1	52.5 ± 3.1
Facial expressions	63.4 ± 4.4	58.8 ± 4.8	61.3 ± 3.5	62.9 ± 2.2
Physiological + Facial expressions	61.6 ± 4.8	60.7 ± 1.5	45.5 ± 3.2	55.4 ± 4.5
Dominance				
Feature Set	**RF**	**XGB**	**KNN**	**SVM**
Physiological	67.5 ± 4.8	67.0 ± 2.7	65.0 ± 3.2	67.1 ± 1.8
Facial expressions	74.3 ± 1.9	75.4 ± 3.0	72.5 ± 3.0	72.5 ± 2.2
Physiological + Facial expressions	73.8 ± 1.8	74.5 ± 0.9	64.1 ± 4.6	65.9 ± 2.8

Table 3.11. SAM classification results showing prediction accuracy of high-/low valence, arousal and dominance values (threshold 4.5 on 0-9 scale) from physiological data, facial expressions, and their combination. Tables report cross-validation mean accuracy (%) and standard deviation. The highest value of accuracy is highlighted.

The classification results indicate that facial expression features outperformed physiological data across all emotional dimensions. However, the fusion of both modalities did not consistently lead to improved classification

performance.

For valence classification, facial expressions achieved the highest accuracy, with 63.4% using Random Forest (RF), 61.3% using k-Nearest Neighbors (KNN), and 62.9% using Support Vector Machines (SVM). The combination of physiological and facial features did not yield a notable improvement over individual modalities.

In arousal classification, the best performance was obtained with facial expressions using RF (63.4%), surpassing both physiological features and their combined use.

For dominance classification, facial expressions achieved the highest accuracy with XGBoost (75.4%), followed by RF (74.3%). Similar to the other emotional dimensions, multimodal fusion did not provide a significant advantage over facial features alone.

Musical Feature Selection and Prediction Our multimodal approach to song classification yielded promising results, with the highest accuracy reaching 67.9% using a combination of video and physiological signals, as well as video and facial expressions (Table 3.12). Notably, different feature sets influenced classifier performance, with XGB and KNN achieving the best scores. These results indicate a significant improvement over previous attempts and highlight the effectiveness of multimodal fusion in tackling this seven-class classification problem.

Feature Set	RF	XGB	KNN	SVM
Video + Physiological	63.6 ± 6.9	67.9 ± 9.3	65.0 ± 10.9	65.7 ± 10.3
Video + Facial expressions	66.4 ± 7.4	62.9 ± 8.6	66.4 ± 10.3	65.0 ± 9.9
Physiological + Facial expressions	36.4 ± 4.2	39.3 ± 11.7	15.7 ± 3.6	24.3 ± 5.7
Video + Physiological + Facial expressions	63.6 ± 6.9	62.9 ± 9.7	67.1 ± 8.9	65.7 ± 10.3

Table 3.12. Audio selection classification results showing prediction accuracy of the selected sound (out of the 7 possible ones) from the combinations of video features, physiological data, and facial expressions. The Table reports cross-validation mean accuracy (%) and standard deviation. The highest value of accuracy is highlighted.

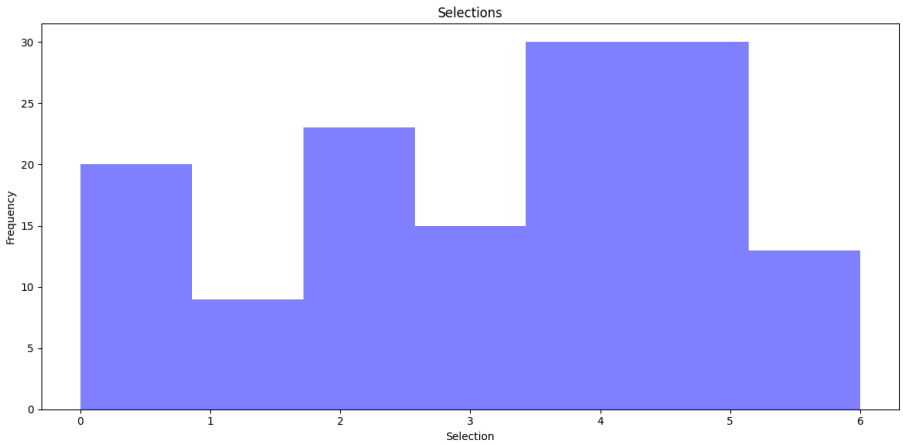


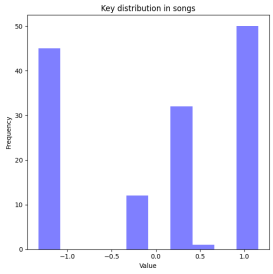
Figure 3.22. Distributions of song selections.

Given these results, we further explored predicting specific musical characteristics. The analysis of the audio stimuli revealed distinct distributions across musical features (Figure 3.23), reinforcing the idea that these individual components might better capture the basis for participants' preferences.

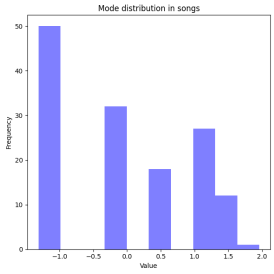
Regression models for predicting musical features showed remarkable performance with certain feature combinations (Table 3.13). The combination of video features with physiological signals achieved exceptionally low error rates (MSE $\approx 10^{-29}$) for most musical characteristics, with Linear Regression and XGBoost showing the best performance. Models relying solely on physiological and facial features showed notably higher error rates, emphasizing the importance of the shown visual content in predicting appropriate musical features.

Discussion

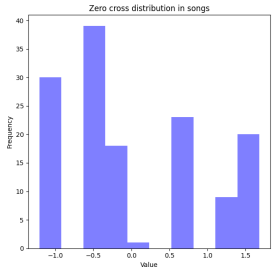
Research Question 1: Valence-Arousal Mapping in Audio-Visual Associations Our findings definitively answer RQ1 by demonstrating that valence dominates arousal in driving audio-visual emotional associations. The strong preference for valence-matched stimuli (83% selection rate) over



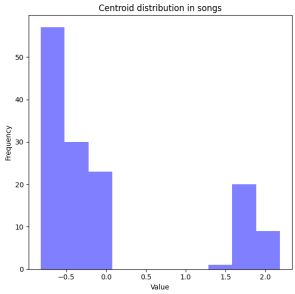
(a) Key distribution



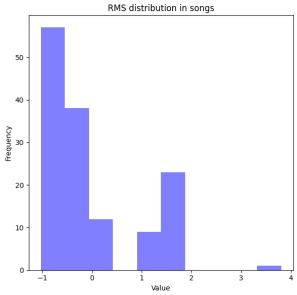
(b) Mode distribution



(c) Zero-crossing rate distribution



(d) Centroid distribution



(e) RMS distribution

Figure 3.23. Distributions of key musical features extracted from the audio stimuli using MIRtoolbox: (a) Shows the key distribution across songs, (b) represents the mode distribution ranging from minor (-1) to major (+1), (c) displays the zero-crossing rate distribution indicating signal complexity, (d) shows the spectral centroid distribution representing the centre of mass of the spectrum, and (e) illustrates the RMS energy distribution showing the signal's amplitude variations.

arousal-matched alternatives (6%) reveals a clear hierarchy in cross-modal emotional processing. This aligns with studies on cross-modal associations [230, 231] while providing novel quantification of the relative importance of these dimensions.

Research Question 2: SAM ratings prediction from psycophysiological responses The predictive relationships between psychophysiological responses and SAM ratings provide valuable insights into the validity of our experimental protocol. The classification results demonstrate a moderate

Key				
Feature Set	**RFR**	**XGBoost**	**SVR**	**LR**
Video + Physiological	0.012 ± 0.022	0.023 ± 0.047	0.222 ± 0.156	1.974 ± 2.901
Video + Facial expressions	0.015 ± 0.029	0.023 ± 0.047	0.417 ± 0.233	1.719 ± 1.700
Physiological + Facial expressions	1.021 ± 0.160	1.077 ± 0.187	1.099 ± 0.123	1.986 ± 0.273
Video + Physiological + Facial expressions	0.012 ± 0.022	0.030 ± 0.044	0.650 ± 0.103	0.723 ± 0.114
Mode				
Feature Set	**RFR**	**XGBoost**	**SVR**	**LR**
Video + Physiological	0.007 ± 0.010	0.014 ± 0.027	0.027 ± 0.017	$(4.82 \pm 2.45) \times 10^{-29}$
Video + Facial expressions	0.006 ± 0.009	0.014 ± 0.027	0.244 ± 0.218	0.153 ± 0.251
Physiological + Facial expressions	1.027 ± 0.104	1.045 ± 0.183	1.086 ± 0.087	1.867 ± 0.237
Video + Physiological + Facial expressions	0.007 ± 0.010	0.015 ± 0.027	0.484 ± 0.037	0.475 ± 0.035
Zero-crossing rate				
Feature Set	**RFR**	**XGBoost**	**SVR**	**LR**
Video + Physiological	0.002 ± 0.002	0.000 ± 0.001	0.037 ± 0.017	$(4.08 \pm 3.85) \times 10^{-29}$
Video + Facial expressions	0.003 ± 0.002	0.000 ± 0.001	0.132 ± 0.037	0.193 ± 0.074
Physiological + Facial expressions	0.728 ± 0.125	0.737 ± 0.119	0.982 ± 0.139	1.503 ± 0.082
Video + Physiological + Facial expressions	0.003 ± 0.003	0.000 ± 0.001	0.277 ± 0.047	0.270 ± 0.050
Centroid				
Feature Set	**RFR**	**XGBoost**	**SVR**	**LR**
Video + Physiological	0.013 ± 0.024	0.026 ± 0.053	0.028 ± 0.010	$(3.71 \pm 3.25) \times 10^{-29}$
Video + Facial expressions	0.011 ± 0.021	0.026 ± 0.053	0.045 ± 0.030	0.042 ± 0.038
Physiological + Facial expressions	0.870 ± 0.144	0.969 ± 0.220	1.129 ± 0.321	2.030 ± 0.123
Video + Physiological + Facial expressions	0.010 ± 0.018	0.027 ± 0.053	0.367 ± 0.117	0.349 ± 0.121
RMS				
Feature Set	**RFR**	**XGBoost**	**SVR**	**LR**
Video + Physiological	0.052 ± 0.089	0.046 ± 0.092	0.029 ± 0.009	$(4.81 \pm 3.82) \times 10^{-29}$
Video + Facial expressions	0.051 ± 0.089	0.046 ± 0.092	0.271 ± 0.202	0.166 ± 0.223
Physiological + Facial expressions	0.995 ± 0.324	1.094 ± 0.421	1.025 ± 0.444	1.689 ± 0.234
Video + Physiological + Facial expressions	0.051 ± 0.087	0.046 ± 0.092	0.516 ± 0.246	0.514 ± 0.222

Table 3.13. Musical features regression results showing prediction mean squared error (MSE) and standard deviation of the computed features of the selected sounds from the combinations of video features, physiological data, and facial expressions. The lowest value of MSE is highlighted.

to strong alignment between objective measurements and subjective self-assessments, particularly for dominance ratings where the facial expressions achieved 75.4% accuracy. This suggests that our protocol successfully captured genuine emotional responses that were consistently reflected in both multimodal behavioural signals and self-reports. The varying accuracy

levels across emotional dimensions (62.9% for valence, 63.4% for arousal, and 75.4% for dominance) indicate that certain aspects of emotional experience may be more reliably captured through psychophysiological measurements than others. This differential performance could be attributed to the inherent complexity of emotional states and the varying degree to which different emotional dimensions manifest in physiological responses.

The consistently strong performance in dominance prediction across all feature combinations (ranging from 67.1% to 75.4%) is particularly noteworthy, as it suggests that our experimental protocol was especially effective in eliciting and measuring this emotional dimension. This finding contributes to the broader understanding of how dominance states manifest in physiological responses and facial expressions. These results ultimately support the validity of our experimental protocol while also providing valuable insights for future refinements in emotion measurement methodology. The demonstrated alignment between objective and subjective measures reinforces the reliability of our findings and provides a solid foundation for addressing subsequent research questions in our study.

Research Question 3: Predicting Musical Features from Emotional Responses Our multimodal approach achieved promising results, with the best model reaching 67.9% accuracy using video and physiological signals, as well as video and facial expressions. These results demonstrate the strong predictive power of visual and physiological features in music preference classification. However, to gain deeper insights into the underlying factors driving these predictions, we shifted focus to musical feature prediction. This approach yielded remarkably precise results, aligning with research highlighting the importance of acoustic characteristics in emotional responses [232].

The exceptional performance of models combining video and physiological features (MSE ranging from 10^{-28} to 10^{-5}) significantly outperformed the combination of video and facial expressions, suggesting that a wearable-based implementation could be more practical and effective than camera-

based solutions. Linear Regression performed particularly well for temporal features (zero-crossing rate, RMS) and spectral characteristics (centroid), while XGBoost excelled at predicting harmonic properties (mode, key). These results suggest different musical characteristics may require distinct modelling approaches for optimal prediction.

A key finding is the critical role of visual content in prediction accuracy. When visual features were removed, song classification accuracy dropped to just 39.3%, and models using only physiological and facial data for musical feature prediction showed substantially higher error rates (MSE > 0.5). This indicates that both tasks—song selection prediction and musical feature regression—require direct access to the visual stimulus rather than relying solely on its emotional impact. These findings have important implications for SSD design, suggesting that optimal performance necessitates both visual processing capabilities and physiological monitoring rather than emotional response measurement alone.

These findings provide a concrete foundation for developing automated sonification systems that maintain emotional coherence. The ability to predict specific musical features with high precision enables granular control over the emotional qualities of generated audio, potentially allowing for more nuanced and engaging sensory substitution experiences.

Conclusions

Our study demonstrates the feasibility of emotion-aware sensory substitution through three key findings. First, valence dominates arousal in audio-visual emotional matching (83% vs 6%), providing clear direction for sensory translation algorithms. Second, our experimental protocol's validity was confirmed by successful SAM prediction from physiological responses (75.4% accuracy for dominance). Finally, the combination of visual features and physiological signals achieves exceptional predictive performance for musical features (MSE 10^{-29}), suggesting a practical path toward wearable implementation.

Implications for Sensory Substitution Design The strong preference for valence-matched stimuli should guide the development of sensory translation algorithms, with priority given to preserving emotional valence over arousal dimensions. The successful multimodal emotional measurement approach, combining physiological sensing with visual processing, indicates that effective SSDs will benefit from integrating multiple input channels. Moreover, the exceptional accuracy in predicting musical features from visual-physiological data provides a concrete foundation for implementing automated, emotion-aware sonification algorithms.

The successful performance achieved using physiological signals rather than facial expressions has significant practical implications, as wearable sensors are generally more suitable for real-world SSD applications than camera-based systems. While MineCLIP-based visual feature extraction currently presents computational challenges for embedded systems, recent advances in model compression [129] and optimisation techniques suggest feasible paths toward practical implementation.

Limitations and Future Directions Several limitations should be considered when interpreting our results. First, our study was conducted in a controlled laboratory environment with a relatively young participant pool (mean age 25.94 years) and gender imbalance (72.2% female). Second, the limited number of video stimuli (4) may not fully represent the range of possible emotional content. Future research should validate the system in real-world environments [233], investigate individual differences in emotional responses across diverse demographic groups [234], and develop lightweight versions of visual feature extraction algorithms suitable for embedded implementation [128]. Additionally, expanding the stimulus set would strengthen the foundation for developing practical, emotion-aware sensory substitution devices that can effectively maintain emotional coherence in audio-visual translation.

3.7 Music Therapy

3.7.1 Introduction

Music therapy is a clinical intervention that uses music experiences to address individuals' physical, emotional, cognitive, and social needs [235]. A qualified music therapist assesses the client's needs and develops a comprehensive treatment plan involving creating, singing, moving to, and/or listening to music [235]. This therapeutic approach systematically harnesses music's inherent power to evoke, express, and regulate emotions in a structured clinical environment.

A growing body of research has demonstrated music's profound influence on emotional states, highlighting its potential as a non-pharmacological intervention for enhancing well-being and quality of life across diverse populations [232, 236]. Neuroimaging studies have shown that listening to music activates brain regions associated with emotion processing and reward pathways, including the amygdala, hippocampus, and nucleus accumbens, leading to measurable physiological changes that correlate with subjective emotional experiences [232]. These physiological responses include alterations in heart rate, blood pressure, respiration, skin conductance, and hormonal secretions, which collectively contribute to the emotional impact of music therapy interventions.

Musical elements like tempo, harmony, melodic structure, dynamic range, and unexpected changes can heighten emotional responses and create therapeutic opportunities [232], while contextual factors such as personal associations, cultural background, setting, and social context shape music's emotional impact and therapeutic efficacy [236]. The flexibility of music as a therapeutic medium allows for personalized interventions that can be tailored to individual preferences, clinical needs, and treatment goals.

Specialized techniques such as songwriting, lyric analysis, the iso principle (matching music to the client's emotional state before gradually modifying it), receptive listening, and clinical improvisation have been shown to ef-

fectively elicit and regulate emotions, increasing positive affect and providing a structured outlet for emotional expression and processing [237, 236, 238]. These evidence-based approaches enable clients to explore, communicate, and transform emotional experiences in ways that may be difficult to access through verbal therapy alone. Additionally, music therapy facilitates beneficial physiological changes that support healing processes and may significantly reduce the need for pharmacological pain management in certain clinical contexts [236].

The therapeutic relationship between the music therapist and client serves as a fundamental element of the intervention, creating a safe space for emotional exploration and expression through musical engagement. This relationship is characterized by empathic attunement, clinical expertise, and collaborative goal-setting that respects the client's autonomy and preferences within the therapeutic process.

This chapter will present three comprehensive music therapy projects carried out by the music therapy team at Università Campus Bio-Medico di Roma in different hospital wards, demonstrating the practical application and emotional benefits of music therapy in diverse clinical settings. These case studies will illustrate how theoretical principles are translated into effective clinical practice, providing valuable insights for healthcare professionals interested in integrating music therapy into multidisciplinary treatment approaches.

3.7.2 Music Therapy Effects on Hemodynamics in Pain Management during Hemodynamic Procedures

Introduction

Cardiac catheterization procedures can induce significant stress responses in patients, potentially affecting their physiological parameters and procedural outcomes. While pharmacological interventions are commonly used to manage patient anxiety, there is growing interest in non-pharmacological approaches such as music therapy. This study investigated the effects of per-

sonalized music therapy on key physiological parameters during diagnostic coronary angiography and percutaneous coronary intervention (PCI) procedures. Previous studies have primarily focused on the psychological benefits of preselected music during cardiac procedures, with limited quantitative analysis of physiological responses. This investigation aimed to address this gap by implementing a comprehensive analysis of vital parameters, including Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Heart Rate (HR), Heart Rate Variability (HRV), and peripheral Oxygen Saturation (SpO2). The study introduced several innovative elements in the experimental design:

- Implementation of real-time shared music listening between patient and music therapist
- Personalization of musical selections based on patient preferences and emotional state
- Continuous monitoring and analysis of physiological parameters throughout the procedure
- Integration of voice intervention by the music therapist during the procedure

Materials and Methods

A prospective study was conducted on 101 patients undergoing cardiac catheterization procedures at the Campus Bio-Medico University Hospital of Rome between June 2022 and May 2023. The study population was divided into an experimental group (n=51) receiving music therapy support and a control group (n=50) undergoing standard procedures, with pharmacological sedation administered as clinically indicated. The protocol of this study was reviewed and approved by the ethical committee of Fondazione Policlinico Universitario Campus Bio-Medico on 1st December 2022 with number of register 2021.232.

Data Acquisition Throughout the entire catheterization procedure, a comprehensive set of physiological parameters was continuously recorded. The monitoring system included seven-lead ECG for continuous cardiac monitoring, invasive arterial blood pressure measurement through an arterial introducer connected to a pressure transducer, and continuous peripheral oxygen saturation monitoring via pulse oximetry. Additional non-invasive blood pressure measurements were taken via sphygmomanometer before and after the procedure. All procedural events were timestamped, including introducer insertion and removal, and contrast medium administration times, to enable subsequent temporal analysis.

Signal Processing and Analysis The analysis protocol involved several stages of data processing. For ECG signals, the highest quality leads were selected for subsequent analysis. Heart Rate Variability was computed using the Root Mean Square of Successive Differences (rMSSD) methodology. From the continuous blood pressure recordings, mean systolic and diastolic pressures were extracted. SpO2 analysis, given its characteristically low variability, utilized representative values from the selected time periods. The raw data underwent quality assessment, leading to the exclusion of measurements with standard deviation exceeding 20% of the relative value for at least 5 measures. Additional exclusions were made based on pre-analytical quality criteria. In the experimental group, patients who required pharmacological intervention were also excluded from the analysis.

Table 3.14. Study Population Distribution After Exclusions

Parameter	Experimental Group	Control Group
Initial Population	51	50
Hemodynamic Analysis	36	36
SpO2 Analysis	38	46

Temporal Analysis From the continuous recordings, four temporal windows were extracted for detailed analysis. The first window was defined from the arterial introducer insertion to the first contrast medium administration, extending slightly beyond in cases where this interval was too brief. The second and third windows were selected from the central portion of the recording based on signal quality optimisation. The fourth window spanned from the final contrast medium administration to introducer removal (Figure 3.24.

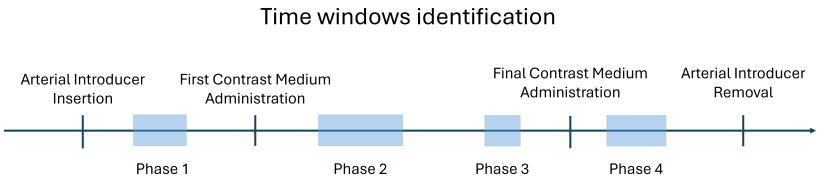


Figure 3.24. Time windows identification during the cardiac catheterization procedure. Four windows were identified based on signal quality assessment: Phase 1 (from arterial introducer insertion to first contrast medium administration), Phases 2 and 3 (selected from the central portion of the recording according to signal quality optimisation), and Phase 4 (from final contrast medium administration to introducer removal). The exact timing and duration of each window was adapted based on data quality criteria.

For each temporal window, mean values and standard deviations were calculated using the first window as reference. Statistical analysis employed ANOVA for between-group comparisons, with a significance threshold of $p \leq 0.05$. The Pearson correlation coefficient was used to assess relationships between variables, with correlations greater than 0.3 in absolute value considered significant.

Musical Parameters Analysis The music therapy sessions were characterized through several musical dimensions. Agogics was measured through beats per minute (BPM), with observed values ranging from 47 to 89.67 BPM. Musical genres encompassed Classical, Light-Pop, Jazz-Latin, New Age-Celtic, Rock, and Film Music compositions. While time signatures included 3/4, 2/4, and 12/8, the majority of pieces were in 4/4 time. Each

piece was classified according to its tonality (Major or Minor), dynamics (mezzo-piano or mezzo-forte), and register (Grave or Medium). Additional parameters included the vocal or instrumental nature of each piece and the presence or absence of music therapist's voice intervention. The analysis methodology examined correlations between these musical parameters and the recorded physiological measurements using Pearson's correlation coefficient. Initial evaluation considered correlations between each musical parameter and the mean physiological values across all temporal windows. Where significant correlations were identified ($|r| > 0.3$), subsequent analysis investigated relationships within individual temporal windows.

Statistical Analysis Continuous variables were expressed as mean and standard error, while discrete variables were presented as number and percentage. Between-group differences in means were evaluated using ANOVA, with statistical significance defined at $p \leq 0.05$. The assessment was conducted in two distinct approaches. The first compared mean physiological parameters between groups within each temporal window. The second analysed relative variations by computing the changes in windows subsequent to the first window, using the first window as reference. For correlation analysis between musical and physiological parameters, the Pearson correlation coefficient was employed. Correlations exceeding 0.3 in absolute value were considered significant. For parameters showing significant overall correlation, additional analysis was performed examining relationships within individual temporal windows. A superimposition test was conducted between the music therapy and control groups to evaluate the reduction in sedation requirements. This analysis employed Wilson confidence intervals and chi-square testing to assess statistical significance. A separate analysis was performed on the subset of patients who underwent percutaneous coronary intervention (PCI) following diagnostic coronary angiography. This subgroup analysis included 6 patients from the experimental group and 18 from the control group, of whom 6 received sedation and 12 did not. The same statistical methodology was applied to this subset,

with SpO2 analysis performed on 6 experimental and 22 control patients (7 sedated, 15 non-sedated).

Results

A superiority test between groups demonstrated a statistically significant reduction in sedation requirements. In the control group, 44% of patients (22/50) required sedation, compared to only 11.8% (6/51) in the music therapy group (CI 95%: -11.6% to -43.3%, p=0.0000097).

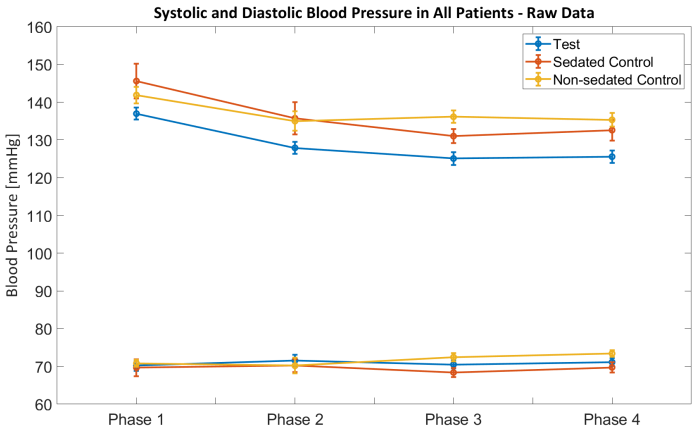
Blood Pressure Analysis Analysis of mean Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) across the four temporal windows revealed no significant differences between experimental and control groups, indicating maintained hemodynamic stability throughout the procedure.

Heart Rate and Heart Rate Variability Analysis Heart Rate (HR) showed significantly higher values in the experimental group compared to controls, particularly evident when compared to sedated patients ($p < 0.05$). The difference was most pronounced in phases 2 and 3 of the procedure.

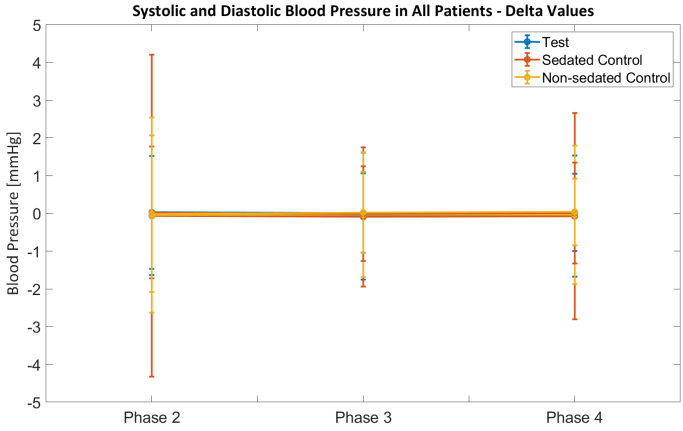
Table 3.15. Statistical Analysis of Heart Rate (p-values from ANOVA test)

Groups Comparison	Phase 1	Phase 2	Phase 3	Phase 4
Sedated vs Non-sedated	0.581	0.721	0.364	0.127
Test vs Sedated	0.038	0.034	0.008	0.012
Test vs Non-sedated	0.109	0.065	0.059	0.120
All three groups	0.073	0.051	0.016	0.024
Test vs Control	0.027	0.016	0.007	0.016

Heart Rate Variability (HRV) analysis showed no significant differences between groups when assessed either as absolute values or as relative changes from baseline.



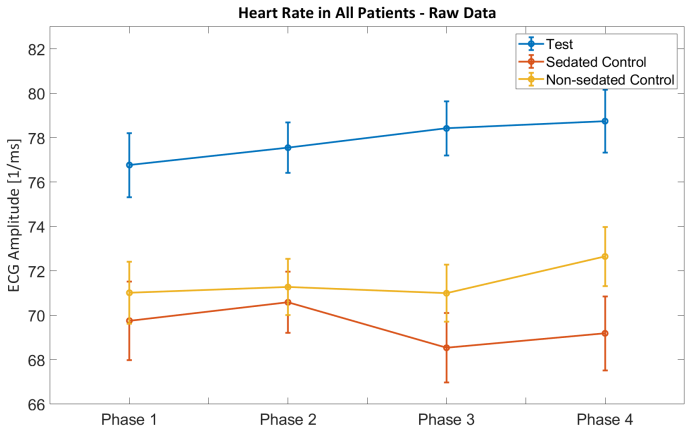
(a) Raw measurements across phases



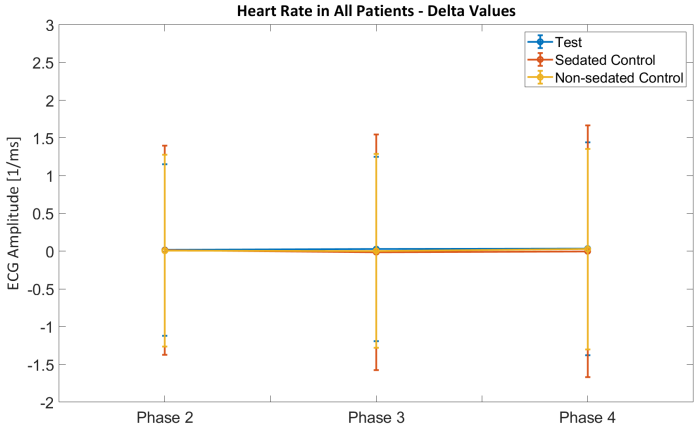
(b) Delta values from baseline

Figure 3.25. Blood Pressure analysis in all patients

Oxygen Saturation Analysis Oxygen saturation (SpO2) analysis revealed significantly higher values in the music therapy group compared to the control group, particularly when compared to non-sedated controls. This improvement in SpO2 was maintained throughout the procedure.



(a) Heart Rate measurements across phases

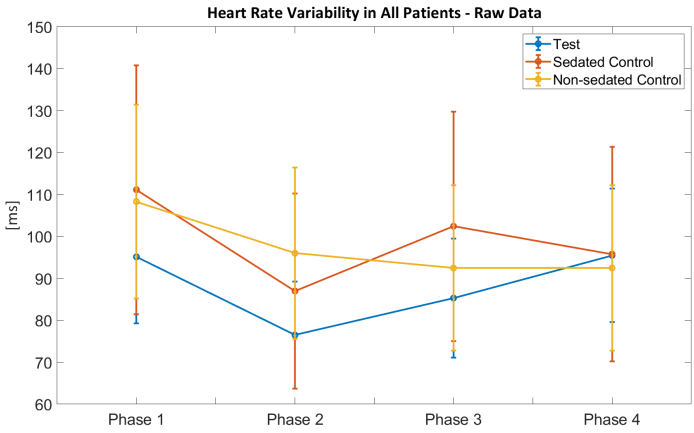


(b) Heart Rate relative changes from baseline

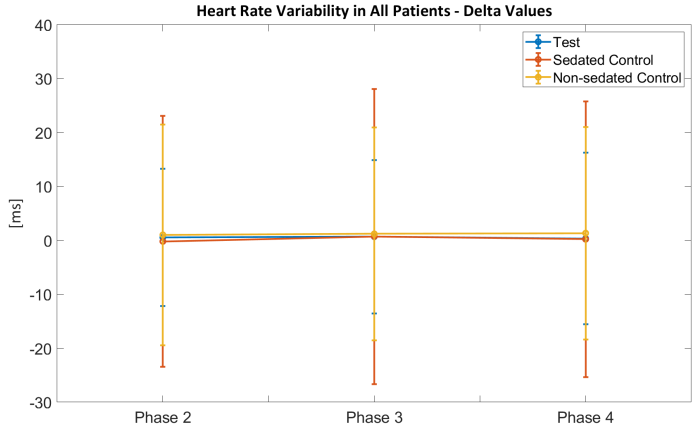
Figure 3.26. Heart Rate analysis in all patients

Angioplasty Subgroup Analysis A subset of patients who underwent percutaneous coronary intervention following diagnostic catheterization was analysed separately (6 experimental vs. 18 control patients, including 6 sedated and 12 non-sedated).

In this subgroup, systolic blood pressure was significantly higher in the experimental group compared to non-sedated controls during phases 1 and



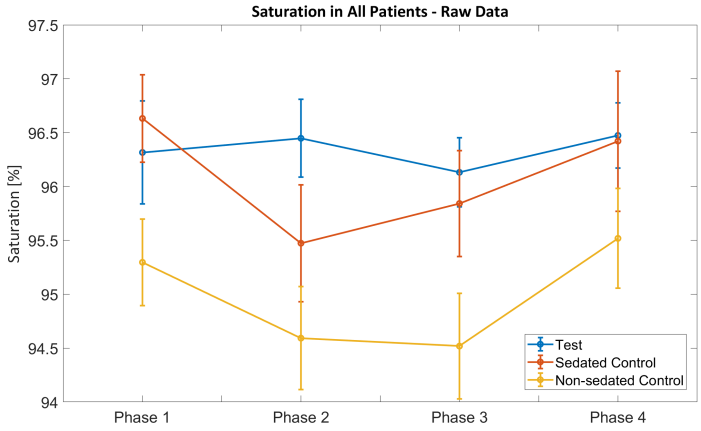
(a) Heart Rate Variability measurements



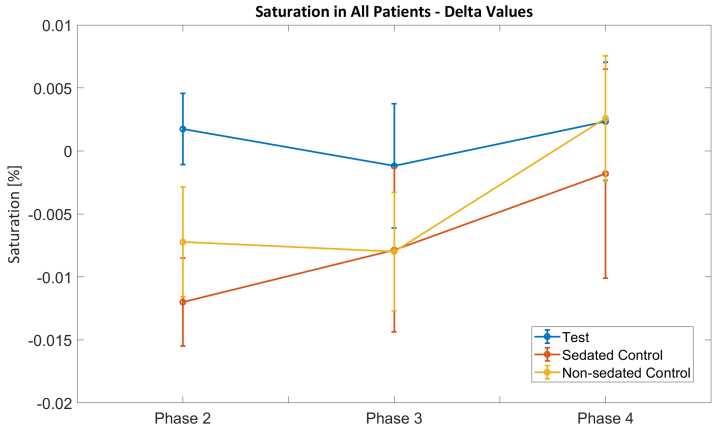
(b) Heart Rate Variability relative changes

Figure 3.27. Heart Rate Variability analysis in all patients

2 ($p < 0.05$). Diastolic blood pressure showed similar patterns, with significant differences in phase 2. Heart rate analysis revealed higher values in the experimental group compared to sedated controls during phases 3 and 4. Oxygen saturation showed significant differences in phase 1 between the experimental group and non-sedated controls.



(a) Oxygen Saturation measurements across phases



(b) Oxygen Saturation relative changes from baseline

Figure 3.28. Oxygen Saturation analysis in all patients

Musical Parameters Correlation Analysis of musical parameters revealed several significant correlations with physiological responses. Music dynamics showed a positive correlation with heart rate (Pearson's r = 0.49), particularly evident during the central phases of the procedure. Beat per minute (BPM) values demonstrated an inverse correlation with HRV (Pearson's r = -0.38), where higher BPM corresponded to lower HRV values. Additionally,

Table 3.16. Statistical Analysis of SpO2 (p-values from ANOVA test)

Groups Comparison	Phase 1	Phase 2	Phase 3	Phase 4
Sedated vs Non-sedated	0.028	0.223	0.080	0.286
Test vs Sedated	0.945	0.115	0.662	0.871
Test vs Non-sedated	0.007	0.002	0.012	0.123
All three groups	0.016	0.008	0.034	0.289
Test vs Control	0.076	0.004	0.053	0.261

musical register settings showed a significant correlation with HRV (Pearson's r = 0.37), with lower registers associated with increased HRV.

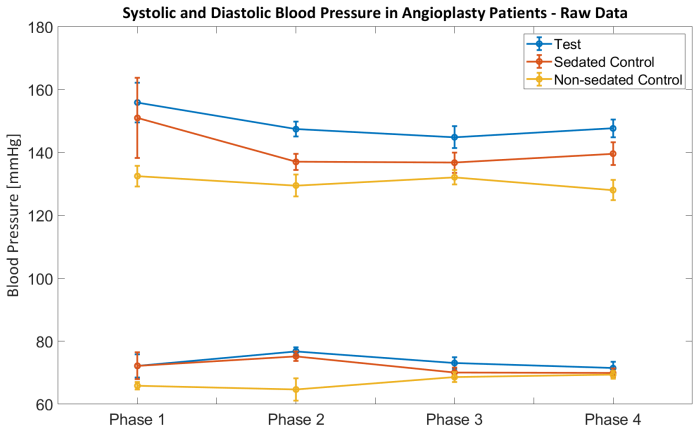
Table 3.17. Correlation Analysis Between Musical and Physiological Parameters

Musical Parameter	Physiological Parameter	Pearson's r
Dynamics	Heart Rate	0.49
BPM	Heart Rate Variability	-0.38
Register	Heart Rate Variability	0.37

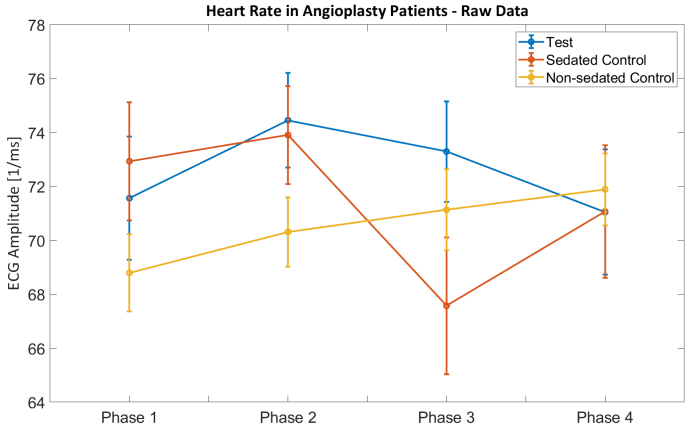
These correlations were most pronounced during phases 2 and 3 of the procedure, suggesting a temporal component to the music therapy effectiveness.

Discussion

The study provides significant evidence supporting the effectiveness of music therapy during cardiac catheterization procedures. The most notable finding was the substantial reduction in sedation requirements in the music therapy group (11.8%) compared to the control group (44%), statistically validated through superiority testing ($p < 0.05$). The analysis of physiological parameters revealed several key findings. Blood pressure measurements showed remarkable stability across all procedural phases in the music



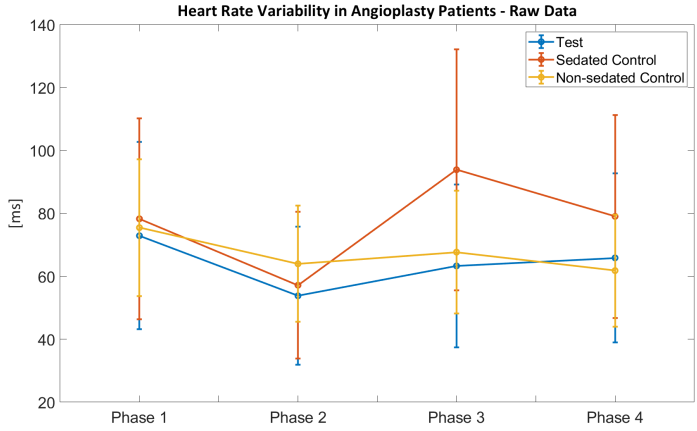
(a) Blood Pressure measurements in angioplasty patients



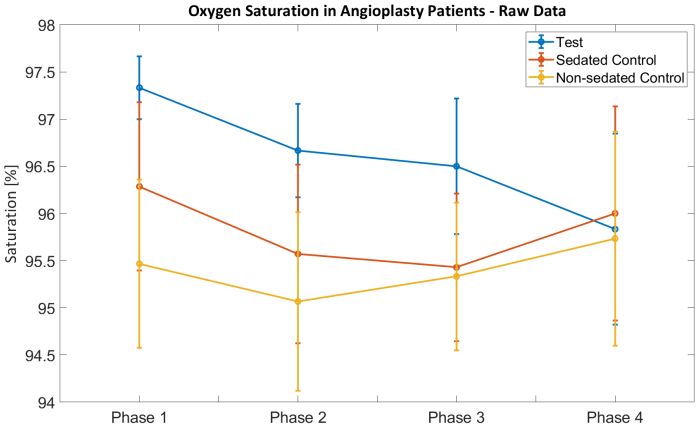
(b) Heart Rate measurements in angioplasty patients

Figure 3.29. Hemodynamic parameters in angioplasty subgroup

therapy group, comparable to controls, indicating effective hemodynamic management without pharmacological intervention. Heart rate exhibited higher values in the music therapy group compared to sedated controls, while maintaining stable heart rate variability. This pattern suggests increased arousal - a state of enhanced cognitive and emotional engagement - rather than stress-induced tachycardia. A particularly significant finding



(a) Heart Rate Variability in angioplasty patients



(b) Oxygen Saturation in angioplasty patients

Figure 3.30. Additional physiological parameters in angioplasty subgroup

was the improved oxygen saturation in the music therapy group compared to non-sedated controls. This enhancement remained stable throughout the procedure, suggesting more efficient respiratory patterns under music therapy. The absence of the typical initial SpO2 decrease observed in sedated patients further supports the benefits of this non-pharmacological approach. Musical parameter analysis revealed meaningful correlations between spe-

cific musical elements and physiological responses. The relationship between music dynamics and heart rate, and between tempo (BPM) and heart rate variability, demonstrates the importance of careful music selection in therapeutic applications. These correlations were most pronounced during the central phases of the procedure, indicating an optimal therapeutic window. In the angioplasty subgroup, despite the limited sample size, observed trends suggest potential benefits of music therapy during more complex procedures. The maintenance of stable hemodynamic parameters without increased sedation requirements is particularly noteworthy in this higher-risk context.

Conclusions

This investigation demonstrates that personalized music therapy represents an effective non-pharmacological support during cardiac catheterization procedures. Key findings include:

- Significant reduction in sedation requirements
- Maintenance of hemodynamic stability
- Enhanced oxygen saturation profiles
- Specific correlations between musical parameters and physiological responses

The study introduces several innovative elements to the field, including real-time shared music listening and continuous physiological monitoring. These advances provide a foundation for further research into music therapy applications in interventional cardiology. Future investigations should focus on expanding the analysis to larger populations, particularly in complex procedures, and developing standardized protocols for music selection based on individual patient characteristics and procedural requirements.

3.7.3 Music in Dementia Assessment: The MiDAS Project

Introduction and Background

The Music in Dementia Assessment Scales (MiDAS) [239] represents a significant advancement in evaluating music therapy's impact on individuals with moderate to advanced dementia. This observational assessment tool measures observable musical engagement in patients with limited verbal capabilities, enabling systematic evaluation of therapeutic outcomes. The assessment protocol implements Visual Analogue Scales (VAS) without anchor points, utilizing 100mm lines labelled from 'none at all' to 'highest', where the maximum score represents each individual's optimal achievable level at their current stage of dementia.

Assessment Methodology

The study implemented a dual-perspective evaluation system involving both music therapist and staff assessments. Music therapists conducted evaluations at the session start and during the most clinically significant 5-minute period, while staff members performed pre-session and post-session assessments. This comprehensive approach allowed for evaluation of both immediate and sustained therapeutic effects. The assessment framework examined five key dimensions: Interest, Response, Initiative, Involvement, and Enjoyment. Each dimension was evaluated on a VAS scale from 0 to 10, providing quantitative metrics for patient engagement and response. Interest assessment focused on attention to activities and environmental stimuli, while Response evaluation measured awareness and interaction capabilities. Initiative tracking captured communication attempts and activity initiation, Involvement measured participation levels, and Enjoyment assessed emotional responses and relaxation states.

Implementation and Data Analysis

The research protocol encompassed 11 patients across 12 sessions, generating a comprehensive dataset of temporal responses to music therapy interventions. Data collection involved four evaluation points per session, creating a robust framework for analysing therapeutic impact.

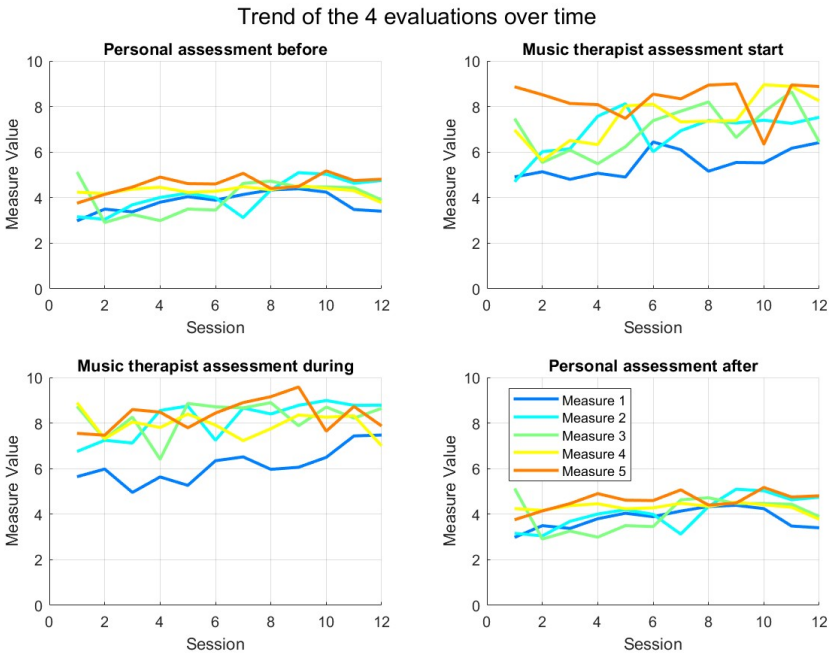


Figure 3.31. Temporal progression of mean values for each measure (Interest, Response, Initiative, Involvement, and Enjoyment) across all evaluation points during the 12-session period. The graph demonstrates overall positive trends with notable improvements in Interest and Involvement metrics.

Temporal analysis revealed consistent improvement patterns across all measures, with particular enhancement in Interest and Involvement metrics. The data demonstrated higher response levels during therapy sessions compared to pre- and post-session evaluations.

Hierarchical clustering analysis using Ward's method identified two distinct patient response patterns, visualized in Figure 3.33. This classification

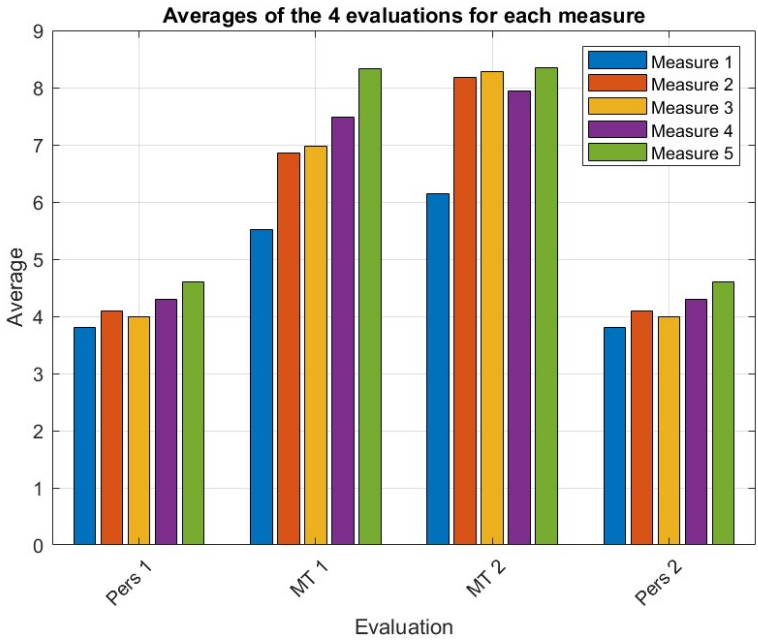


Figure 3.32. Comparative analysis of mean scores across assessment measures for all four evaluation points. Higher scores are evident during music therapy sessions, particularly in Involvement and Enjoyment dimensions.

enabled more targeted therapeutic approaches based on response characteristics.

Clinical Implications and Future Directions

The analysis revealed significant implications for clinical practice, particularly in therapeutic approach customization. Strong responders (Cluster 1) demonstrated sustained improvement and higher baseline scores, while moderate responders (Cluster 2) showed more variable outcomes, suggesting the need for adapted intervention strategies. The study's limitations, including the modest sample size and data imputation requirements, indicate opportunities for future research expansion. Recommendations include larger-scale validation studies, development of supplementary assessment tools, and investigation of long-term benefit retention. The integra-

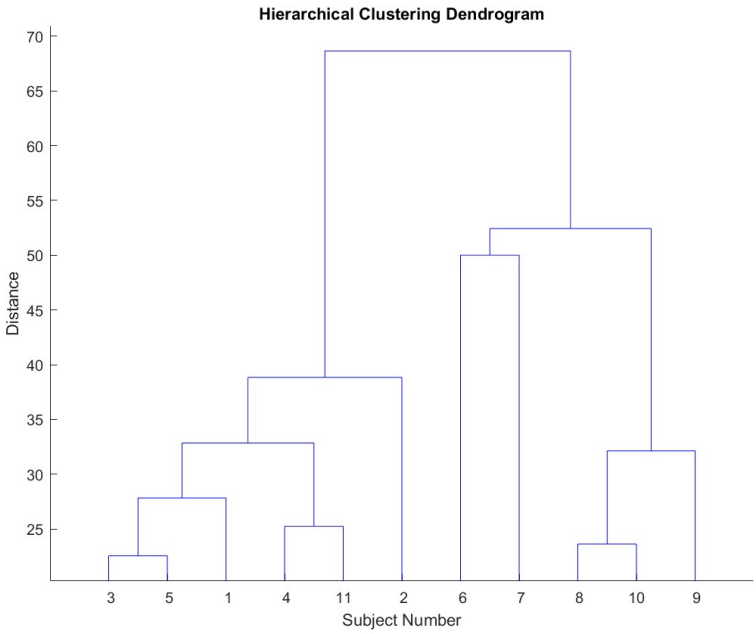
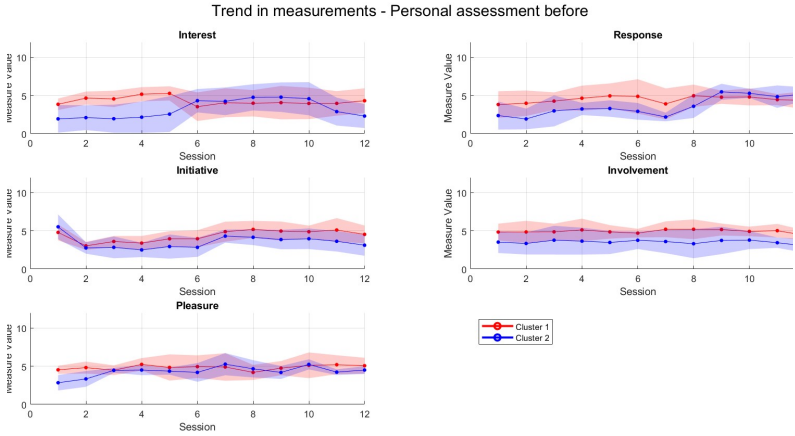


Figure 3.33. Hierarchical clustering dendrogram illustrating patient response patterns, identifying two distinct clusters with varying therapeutic response characteristics.

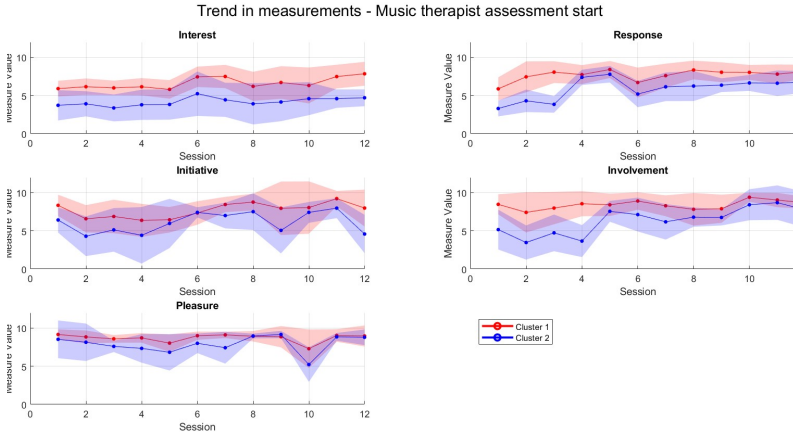
tion of qualitative assessment methods could provide additional insights into therapeutic outcomes. This pilot study establishes a foundation for continued development in music therapy assessment for dementia patients, emphasizing the importance of individualized approaches and systematic monitoring in therapeutic interventions. The findings support the feasibility of structured assessment in music therapy while highlighting areas for methodology refinement and protocol enhancement.

3.7.4 Music Therapy in Cardiothoracic Surgery

This section presents a study examining music therapy effects on cardiothoracic surgery patients through analysis of physiological parameters. The investigation involved 99 patients, with 59 receiving music therapy and



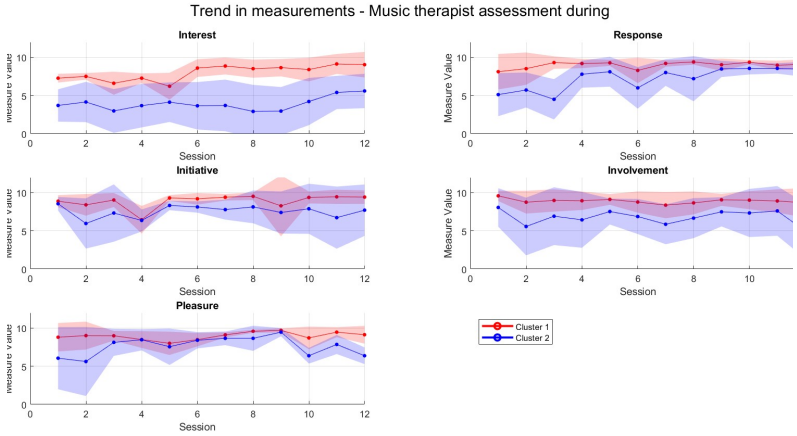
(a) Staff Assessment Before Sessions



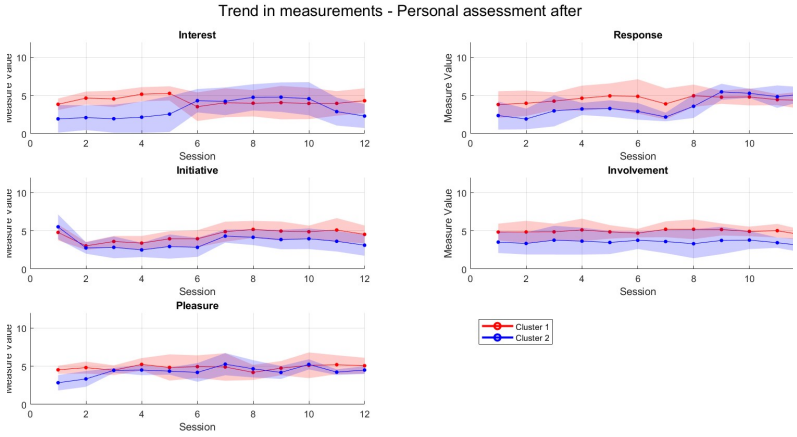
(b) Music Therapist Assessment During Sessions

Figure 3.34. Trends in Music Therapy Assessments (Part 1). Each plot shows measurements of Interest, Response, Initiative, Involvement, and Pleasure for two clusters of patients (Cluster 1 in red, Cluster 2 in blue). Shaded areas represent confidence intervals.

40 serving as controls.



(a) Music Therapist Assessment After Sessions



(b) Staff Assessment After Sessions

Figure 3.35. Trends in Music Therapy Assessments (Part 2). Each plot shows measurements of Interest, Response, Initiative, Involvement, and Pleasure for two clusters of patients (Cluster 1 in red, Cluster 2 in blue). Shaded areas represent confidence intervals.

Study Design and Data Collection Patient monitoring included respiratory support assessment through nasal cannula, Venturi mask, non-rebreather mask, and high-flow oxygen therapy. Cardiac parameters encompassed

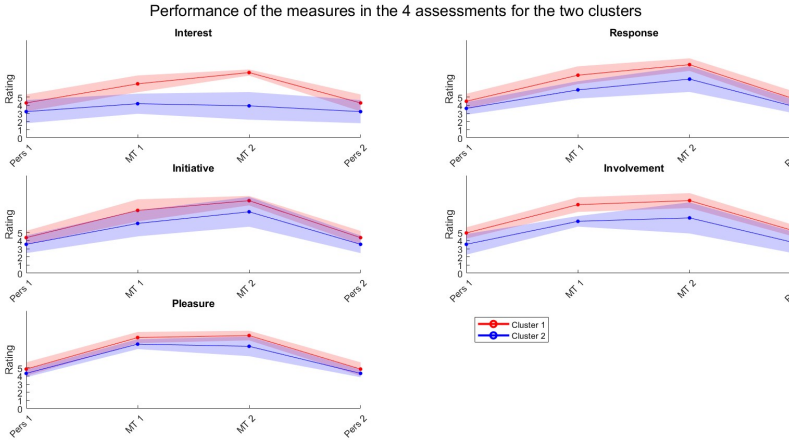


Figure 3.36. Comprehensive overview of cluster differences across evaluation points for each measured dimension, illustrating distinct response characteristics and therapy effectiveness patterns.

sinus rhythm and pacemaker presence. Measurements occurred at rest, during therapy, and post-intervention, tracking systolic and diastolic blood pressure, heart rate, oxygen saturation, and respiratory rate.

Methodology Data preprocessing involved normalization relative to resting values using:

$$\text{normalized value} = \frac{raw\_value - resting\_value}{resting\_value} \tag{3.3}$$

Patient stratification considered oxygen support requirements and therapy exposure. Statistical analysis employed Lilliefors test for normality (p-value = 0.05), followed by ANOVA or Kruskal-Wallis tests as appropriate.

Results

Analysis revealed distinct patterns across patient groups. Patients without oxygen support showed no significant variations in physiological parameters. However, oxygen-dependent patients demonstrated a 10% decrease

Table 3.18. Study Population Distribution

Group	Size
Music therapy - Total	59
Without oxygen support	53
With oxygen support	6
Control - Total	40
Without oxygen support	40
With oxygen support	0
Music therapy - Cluster 1	32
Music therapy - Cluster 2	18

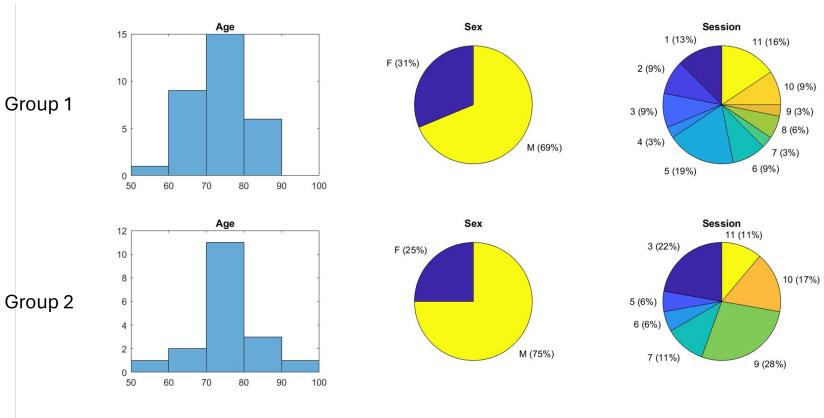


Figure 3.37. Distribution characteristics of physiological parameters in Clusters 1 and 2

in systolic pressure during therapy and 2% increase in oxygen saturation post-therapy.

Cluster analysis identified two distinct response patterns. Cluster 1 exhibited a 2.5% decrease in systolic pressure, 2% decrease in diastolic pressure, and 0.3% increase in heart rate during therapy. Cluster 2 showed similar pressure decreases with additional minor oxygen saturation changes.

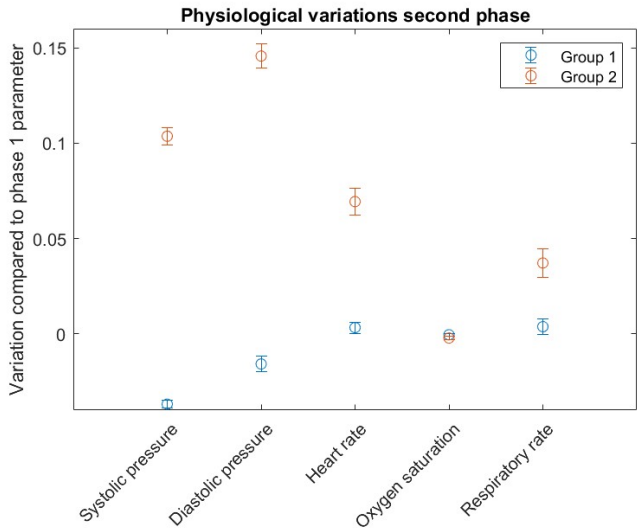


Figure 3.38. Physiological parameter variations during therapy phase for Clusters 1 and 2

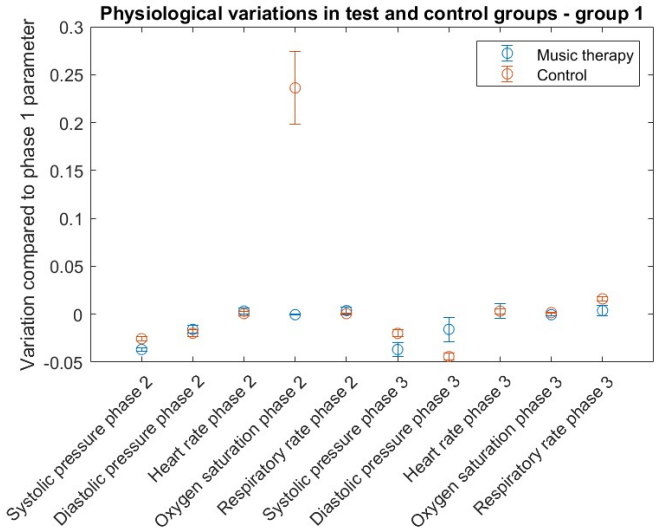


Figure 3.39. Physiological parameter variations across therapy phases in Cluster 1 compared to control group

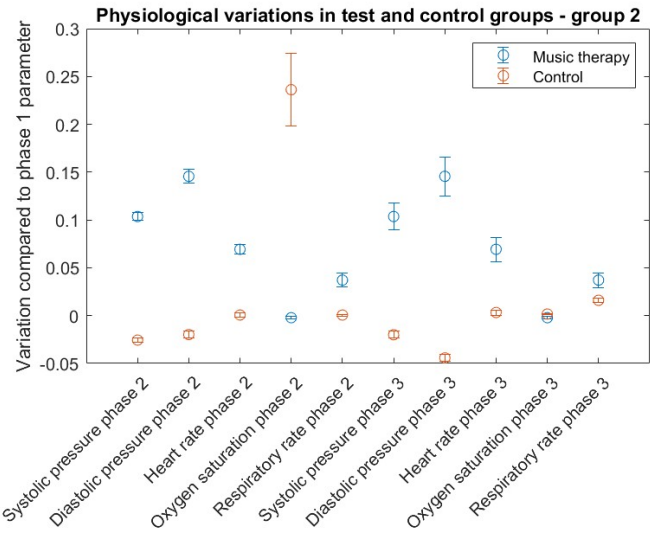


Figure 3.40. Physiological parameter variations across therapy phases in Cluster 2 compared to control group

Discussion

The investigation revealed varying responses to music therapy across patient groups. Non-oxygen-dependent patients maintained stable parameters, suggesting either limited physiological impact or beneficial homeostatic maintenance. Oxygen-dependent patients demonstrated potentially therapeutic responses through reduced systolic pressure and improved saturation.

Conclusions

Music therapy demonstrated measurable physiological effects in cardiothoracic surgery patients, particularly those requiring respiratory support. The varied responses between groups emphasize the importance of individual patient characteristics in therapy outcomes. Future research should address larger sample sizes, extended follow-up periods, and mechanism investigation.

3.7.5 Music Therapy and Affective Computing: Connections to Human-Centred Technology Design

The music therapy applications presented in this chapter illustrate the practical implementation of the theoretical frameworks discussed throughout this thesis, particularly demonstrating how affective computing can enhance human-centred approaches in healthcare contexts. These case studies embody the core principles outlined in Chapter 1, where we emphasized the development of unobtrusive technological solutions that enhance quality of life while maintaining natural human interaction patterns.

The case studies presented in this section demonstrate the practical realisation of the foundational objectives of this research: understanding human perception and emotion, developing emotion-aware technologies, and advancing non-invasive sensing solutions. While current music therapy practice relies primarily on the therapist's observational skills and clinical experience, the integration of affective computing techniques offers several promising avenues for advancing this field while adhering to the human-centred design principles that guide this thesis.

The implementation of real-time emotion recognition systems, as illustrated in the cardiac catheterization study, exemplifies the practical application of the non-intrusive sensing approaches discussed in Chapter 1. By monitoring physiological parameters during music therapy interventions, we were able to quantify therapeutic benefits while minimizing disruption to the patient experience—directly addressing our research objective of developing sensing approaches that minimize user intervention. Similarly, the MiDAS project demonstrates how structured observational tools can complement technological approaches in emotion assessment, creating a more comprehensive understanding of emotional responses in individuals with limited expressive capabilities.

These applications directly connect to the crossmodal perception research presented in Chapter 2, as music therapy inherently leverages the relationship between auditory stimuli and emotional responses. The im-

provement in oxygen saturation observed in cardiac patients and the enhanced engagement noted in dementia patients highlight how emotionally appropriate sensory experiences can positively impact physiological and behavioural responses, reinforcing our findings on emotional mediation in crossmodal perception.

From a human-centred design perspective, these music therapy applications demonstrate several key principles outlined in our research framework. First, they prioritize user comfort and natural experience over technological intrusion, as evidenced by the reduced need for pharmacological interventions in the cardiac catheterization study. Second, they accommodate individual differences in response patterns, as demonstrated by the cluster analysis in both the dementia and cardiothoracic surgery studies. Third, they integrate emotional awareness into therapeutic interventions, aligning with our objective of creating systems that adapt to and work with human emotional processes.

Several practical considerations must be addressed for successful implementation in clinical environments. Non-intrusive sensing technologies are essential to avoid disrupting the therapeutic relationship, with wearable devices offering particular promise due to their minimal interference—a direct application of the low-impact design approach advocated throughout this thesis. The processing of emotional data must occur in real-time to support therapeutic decision-making, requiring systems that balance computational efficiency with clinical accuracy. Additionally, user interfaces must be designed specifically for the therapy setting, providing relevant information without distracting from patient interaction, reflecting the human-centred design principles that inform our work.

The projects described here represent initial steps toward this integration, demonstrating how the theoretical frameworks of emotion recognition, physiological monitoring, and human-centred design can enhance clinical practice in music therapy. Future developments might include adaptive systems that suggest musical interventions based on detected emotional states, longitudinal tracking tools that monitor emotional changes across

multiple sessions, and multimodal platforms that integrate various emotional assessment techniques for comprehensive evaluation—all guided by the principle that technology should enhance human experience while minimizing its impact on natural behaviour.

By combining the humanistic foundations of music therapy with the analytical capabilities of affective computing, these studies exemplify our broader research goal: developing technologies that enhance human capabilities while minimizing their impact on natural behaviour and experience. They demonstrate that when technological solutions are designed with human emotional and perceptual processes at their centre, they can achieve better outcomes while maintaining the natural flow of human interaction.

Chapter 4

Sensing Technology and Calibration

Gian Lorenzo Bernini, "Apollo and Daphne" (1622-1625), marble sculpture, 243 cm height, Galleria Borghese, Rome.

4.1 Introduction to Sensing Technology in Healthcare

The landscape of healthcare monitoring and diagnostics has undergone a transformative evolution, shifting from traditional invasive methodologies towards continuous, non-invasive approaches that prioritise both clinical efficacy and patient experience [240]. This paradigm shift has been driven by technological advancements in sensing technologies, data analytics, and artificial intelligence, enabling unprecedented capabilities in continuous health monitoring whilst minimising patient discomfort and intervention requirements.

4.1.1 Evolution of Healthcare Sensing Technologies

The trajectory of healthcare sensing technologies reflects a fundamental reconceptualisation of patient monitoring approaches. Traditional methodologies, characterised by discrete measurements and often invasive procedures, have increasingly given way to continuous monitoring systems that offer real-time insights into patient health status [241]. This evolution has been particularly pronounced in the management of chronic conditions, where continuous monitoring can provide early warning signs of deterioration and enable more timely interventions [242]. The proliferation of wearable technologies has been a crucial driver in this transformation. Contemporary wearable devices incorporate sophisticated sensor arrays capable of monitoring multiple physiological parameters simultaneously, from basic vital signs to complex biomarkers. These developments have facilitated a shift from reactive to proactive healthcare management, enabling predictive analytics and early intervention strategies [240]. Market analyses indicate substantial growth in the remote patient monitoring (RPM) sector, with projections suggesting a compound annual growth rate of 8.74% through 2030. This growth reflects both technological advancement and increasing recognition of the value of continuous monitoring in improving

patient outcomes and reducing healthcare costs [241].

4.1.2 Human-Centred Design Principles in Healthcare Sensing

The effectiveness of healthcare sensing technologies is intrinsically linked to their integration into patients' daily lives and routines. User-centred design principles have emerged as critical factors in the development and implementation of these technologies [243]. The challenge lies not merely in creating technically capable devices, but in ensuring they are accessible, comfortable, and minimally disruptive to normal activities. Research has demonstrated that sensor design significantly influences patient compliance and adherence to monitoring protocols [244]. Successful implementation requires careful consideration of various factors:

- Ergonomic design and physical comfort during prolonged use
- Intuitive user interfaces that minimise cognitive load
- Clear and actionable feedback mechanisms
- Robust data privacy and security measures
- Integration with existing healthcare workflows

Recent advances in materials science and miniaturisation have enabled the development of increasingly unobtrusive sensing solutions. For instance, next-generation wearable sensors utilising flexible electronics and smart textiles have demonstrated enhanced user acceptance while maintaining measurement accuracy [245]. These developments represent significant progress toward truly non-invasive monitoring solutions that can seamlessly integrate into patients' lives.

4.1.3 Current Technological Landscape

The contemporary healthcare sensing ecosystem is characterised by the convergence of multiple technological advances, particularly in the domains of sensor miniaturisation, artificial intelligence, and Internet of Things (IoT) integration. This convergence has enabled unprecedented capabilities in continuous health monitoring and real-time data analytics [246].

Advanced Sensing Modalities

Modern healthcare sensors employ diverse sensing modalities, each offering specific advantages for different monitoring applications:

- **Electrochemical Sensors:** Utilised in continuous glucose monitoring and blood chemistry analysis
- **Optical Sensors:** Applied in pulse oximetry, photoplethysmography, and spectroscopic analysis
- **MEMS-based Sensors:** Employed in motion detection, respiratory monitoring, and cardiovascular assessment
- **Biochemical Sensors:** Used for monitoring biomarkers in various bodily fluids

The integration of these sensing modalities with artificial intelligence has significantly enhanced their capabilities. Machine learning algorithms can now process complex sensor data in real-time, enabling more accurate detection of physiological anomalies and reduction of false alarms [247]. This integration has been particularly transformative in applications requiring continuous monitoring and rapid response, such as cardiac monitoring and glucose level management.

Data Analytics and Decision Support

Advanced analytics platforms have become integral components of modern healthcare sensing systems. These platforms incorporate:

1. Real-time data processing and feature extraction
2. Machine learning-based pattern recognition
3. Predictive analytics for early warning systems
4. Automated decision support algorithms

The implementation of these analytical capabilities has enabled more sophisticated approaches to patient monitoring. For instance, modern systems can now detect subtle patterns in physiological parameters that may indicate impending clinical events, allowing for preventive interventions [248].

4.1.4 Technical and Human Challenges

The advancement of healthcare sensing technologies faces several significant challenges that span both technical limitations and human factors. From a technical perspective, sensor reliability and accuracy remain paramount concerns in healthcare applications. The achievement of stable, precise measurements in continuous monitoring scenarios is complicated by various factors including signal drift, environmental interference, and motion artifacts. These issues are particularly pronounced in wearable devices, where the dynamic nature of real-world usage introduces additional complexities to the measurement process [249]. Sensor drift represents a particularly challenging aspect of continuous monitoring systems. Over time, sensor performance can degrade due to various factors including material ageing, biofouling, and environmental exposure. This degradation necessitates regular calibration procedures, which must be carefully balanced against user convenience and compliance considerations. The development of robust drift compensation strategies remains an active area of research, with particular emphasis on reducing calibration frequency while maintaining measurement accuracy [250]. Power management presents another critical technical challenge, especially in portable and wearable devices. The requirement for continuous operation must be balanced against battery

life and device size constraints. While advances in low-power electronics and energy harvesting technologies offer promising solutions, the integration of these approaches while maintaining measurement performance requires careful optimisation of system architecture and operation protocols. Beyond technical considerations, the successful implementation of healthcare sensing technologies depends heavily on their integration into existing healthcare systems and workflows. This integration must address not only technical compatibility but also the human factors that influence technology adoption and utilisation. Healthcare providers must be trained in the interpretation of continuous monitoring data, and systems must be designed to present information in a manner that facilitates clinical decision-making without contributing to information overload [251].

4.1.5 Emerging Opportunities

The convergence of advanced sensing technologies with artificial intelligence and data analytics presents transformative opportunities in healthcare monitoring. Artificial intelligence is enabling increasingly sophisticated approaches to personalised healthcare monitoring, with systems capable of adapting to individual patient characteristics and circumstances [252]. These adaptive systems can learn from patient-specific data patterns, adjusting monitoring parameters and alarm thresholds to optimise both clinical utility and user experience. The integration of multiple sensing modalities represents another significant opportunity for advancement in healthcare monitoring. By combining data from complementary sensors, systems can achieve more robust measurements and comprehensive physiological monitoring. This multi-modal approach enables cross-validation between different measurement techniques, enhancing reliability while potentially reducing the frequency of calibration requirements [253]. Advances in materials science and manufacturing technologies are enabling the development of increasingly sophisticated sensor platforms. Novel materials and fabrication techniques allow for the creation of flexible, biocompatible

sensors that can conform to body contours while maintaining measurement accuracy. These developments are particularly relevant for long-term monitoring applications, where user comfort and device durability are critical considerations.

4.2 Research Objectives and Chapter Overview

This thesis addresses fundamental challenges in healthcare sensing technology through investigations in three interconnected areas. In the domain of glucose sensor calibration, we explore novel approaches to drift compensation and calibration optimisation, with particular emphasis on reducing user intervention requirements while maintaining measurement accuracy. This work encompasses both algorithmic developments in signal processing and practical implementations in continuous glucose monitoring systems.

The investigation of breath analysis technologies presents unique opportunities for non-invasive disease monitoring. Our research in this area, published on Sensors and Actuators: B. Chemical [254], focuses on the development and validation of advanced sensor arrays for volatile organic compound detection, with particular emphasis on standardization of measurement protocols and integration with clinical decision support systems. This work addresses both the technical challenges of sensor design and the practical considerations of clinical implementation.

In the field of spectrophotometric analysis, our research explores the development of compact, efficient systems for non-invasive diagnostics. This work, presented with a poster at the European Association of Urology in Paris, France, in 2024, encompasses both hardware optimisation for portable spectrometers and the development of sophisticated algorithms for spectral data analysis. The integration of machine learning techniques enables enhanced pattern recognition and classification capabilities, facilitating rapid and accurate diagnostic assessments.

These research areas share common themes in the pursuit of non-invasive, reliable monitoring solutions that can be integrated into clinical practice. The subsequent chapters present detailed investigations into each of these areas, including theoretical foundations, experimental methodologies, and clinical validation studies. Through these investigations, this thesis contributes to the advancement of healthcare sensing technology by addressing fundamental challenges in sensor development and implementation.

4.3 Glucose Sensor Calibration

4.3.1 Introduction

Diabetes represents one of the most formidable global health challenges of our time, characterized by chronically elevated blood glucose levels that can lead to severe complications including cardiovascular disease, kidney dysfunction, retinal damage, and neuropathy [255]. Recent estimates reveal a staggering statistic: approximately 422 million people worldwide suffer from this metabolic disorder, with projections indicating a surge to 578 million affected individuals by 2030 [256]. This epidemic is particularly concerning in low- and middle-income countries, where approximately 79% of adults with diabetes reside [257].

Glucose Monitoring Technologies

The cornerstone of diabetes management lies in maintaining blood glucose levels within a narrow therapeutic range, typically between 70 and 180 mg/dL [258]. This precise control is primarily achieved through insulin therapy, whose effectiveness critically depends on accurate and timely glucose monitoring. Two primary approaches exist for monitoring blood glucose levels: Blood Glucose Monitoring (BGM) and Continuous Glucose Monitoring (CGM) [259].

Traditional BGM methods, while providing precise readings through finger-prick devices, present significant limitations. The invasive nature of repeated finger-pricking not only causes patient discomfort but also provides only discrete measurements, making it challenging to identify important glucose trends and patterns [260]. These limitations have driven the development of CGM systems.

CGM technology represents a significant advancement in diabetes care, offering continuous measurement of glucose levels in the interstitial fluid through minimally invasive sensors. These systems provide real-time glucose readings and trend information, enabling more proactive diabetes man-

agement [261]. However, CGM technology faces several critical challenges that impact its widespread adoption and effectiveness.

Challenges in CGM Technology

The accuracy of CGM systems is influenced by multiple factors that can affect sensor performance and reliability. These include:

- Physiological variations, such as the lag time between blood and interstitial glucose levels
- Environmental conditions, including temperature and humidity
- Sensor degradation over time
- Calibration accuracy and frequency
- Individual patient characteristics

Current commercial CGM sensors typically employ factory calibration, often supplemented by periodic user calibration through traditional blood glucose measurements. However, the complexity of interfering factors makes maintaining consistent accuracy challenging throughout the sensor's lifetime [262]. Understanding and compensating for these various influences represents a critical challenge in improving CGM performance.

Recent Advances and Current Limitations

Recent technological advances, particularly in the fields of machine learning and edge computing, have opened new possibilities for enhancing CGM accuracy. These technologies enable more sophisticated calibration approaches that can adapt to individual patient characteristics and environmental conditions. However, implementing such solutions on resource-constrained devices presents additional challenges, requiring careful consideration of computational efficiency and power consumption [263].

The development of effective calibration strategies requires a deep understanding of how various factors affect sensor performance, along with the ability to generate realistic data for algorithm development and testing. Furthermore, any proposed solution must be computationally efficient enough to run on the limited hardware typically available in CGM devices.

Research Focus Area

This thesis presents three interconnected contributions addressing the fundamental challenges in CGM sensor calibration:

First, we present a comprehensive analysis of various interfering factors affecting sensor accuracy, including physiological changes, environmental conditions, and sensor-specific characteristics. This work, presented during the 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) [264], provides a mathematical framework for understanding and modelling the impact of these factors on sensor performance.

Building on this understanding, we then introduce a novel approach for generating synthetic but realistic sensor response data. This model, published in BioMedInformatics [265], incorporates both physiological variables and sensor characteristics, providing a robust foundation for developing and testing calibration algorithms. The synthetic data generation approach enables the exploration of various scenarios and conditions that might be difficult or impractical to reproduce in clinical settings.

Finally, we investigate the implementation of lightweight neural network architectures for real-time sensor calibration. This work, published in IEEE Sensors Letters and presented during IEEE Sensors Conference 2024 in Kobe, Japan [**cenerini2024optimising**], focuses on optimising both accuracy and computational efficiency, exploring various neural network architectures and their suitability for deployment on resource-constrained devices.

Through these contributions, we aim to advance the field of CGM technology by improving sensor calibration methods and ultimately enhancing the quality of diabetes care. Our work combines theoretical modelling with

practical implementation considerations, focusing on solutions that can be deployed in real-world devices while maintaining the stringent accuracy requirements necessary for effective diabetes management.

4.3.2 Study on the Impact of Interfering Factors on a Glucose Sensor Model

Background and Research Questions

Over the years, a series of studies have focused on modeling the inherent errors associated with GCM sensors, shedding light on critical aspects that impact their accuracy and reliability. In the early 2010s, Krouwer and Cembrowski emphasized the need for standards and statistics to comprehensively describe blood glucose monitor performance [266]. This laid the groundwork for a holistic approach to evaluating CGM sensors, considering not only analytical errors but also addressing medical errors that could potentially harm patients. Building on this foundation, Facchinetti *et al.* delved into modelling the glucose sensor error [267]. Their work highlighted the challenges faced by CGM sensors, citing distortions due to diffusion processes, time-varying systematic under/overestimations from calibrations and sensor drifts, and the presence of measurement noise. This study underscored the importance of a reliable model for CGM inaccuracies in various applications, such as designing optimal digital filters, real-time glucose prediction, and developing algorithms for artificial pancreas control.

The evolution of CGM sensor technology is evident in subsequent studies, particularly those focusing on Dexcom sensors. In another study, Facchinetti identified and evaluated error models for the G4 Platinum (G4P) and advanced G4 for artificial pancreas studies [268]. The study demonstrated technological advancements, with G4P outperforming its predecessor, the SEVEN Plus, and G4AP showcasing further reliability due to sophisticated data processing algorithms. Vettoretti *et al.* expanded the scope to self-monitoring blood glucose (SMBG) measurements, proposing a novel method-

ology to derive more realistic models of SMBG error probability density functions (PDFs) [269]. The study divided the blood glucose range into zones, each characterized by a constant standard deviation. This innovative approach addressed the limitations of traditional Gaussian models, providing more accurate representations of experimental data.

The advent of factory-calibrated CGM sensors, exemplified by Dexcom G6, prompted further investigation. Vettoretti developed a model dissecting the error into BG-to-IG kinetics, calibration error, and measurement noise [270]. This model extended the applicability to the entire sensor lifetime, a significant advancement considering the 10-day duration of these sensors. In the realm of long-term glucose forecasting, Liu *et al.* proposed an algorithm based on physiological models and deconvolution of CGM signals [271]. Their work addressed the challenge of accurate long-term predictions, a crucial aspect for applications such as precision insulin dosing and artificial pancreas systems.

This work aims to examine the drift or measurement error in CGM sensors that may arise from a variety of interfering variables. These variables, which can be attributed to both external and internal factors, include the patient's behaviour, environmental fluctuations, and the intrinsic characteristics of the sensors themselves. Each effect is modelled and analysed individually. In detail, the focus is on:

- Alterations in the subject's physiological state, induced by sweating or physical exertion, correlate with fluctuations in the rate of glucose change or variations in pH levels;

- Exposure of the sensor to elevated temperatures: this can lead to potential malfunctions or deviations in sensor performance due to thermal stress;

- Compromised adhesion of the sensor: the sensor may experience a reduction in its adhesive properties, which can result in detachment or positional instability.

- Sensor ageing, induced by exogenous skin response, or sensor damage during use

Sources of error and distortion

A CGM system is made up of subcomponents and systems that transform biological information into shareable and meaningful information. In that system, three layers can be found: physical, analogue, and digital layer as shown in Figure 4.1. The glucose concentration in blood, that is the information of interest, must be transduced into a signal that can be acquired and processed. Every sensor must present this typical structure in which the first layer is composed of a transduction layer that can transform a physical variation in analogue information (electrical current/ voltage). In the case of CGM sensors, the glucose transducer can be an electrochemical sensor with glucose oxidase, or an optical sensor in the fluorescent domain or a glucose-binding polymer. The interface circuitry depends on the design choices and can be used as transducer interfaces a current mirror, a trans-impedance amplifier, a charge integrator or a switched capacitor circuit [272], then the signal must be conditioned (with an amplification block and a filtering stage) to enhance signal quality and optimise it for quantization. At this stage, the electrical signal is converted into a digital form and can be transmitted to other systems.

Distorting Factor / Interfering Variables Corruption noise or errors can potentially impact every component within the measurement chain, leading to false alarms or incorrect sensor readings. These measurement inaccuracies may originate from the sensors themselves, the subject being monitored, or environmental factors. Equation 4.1 describes the measured value, $y(t)$, as a combination of the true value, $u(t)$, affected by distortion and noise, capturing the realistic behaviour of the measurement system. In the equation appears also the time dependency that affects the true value, the distortion function and the noise.

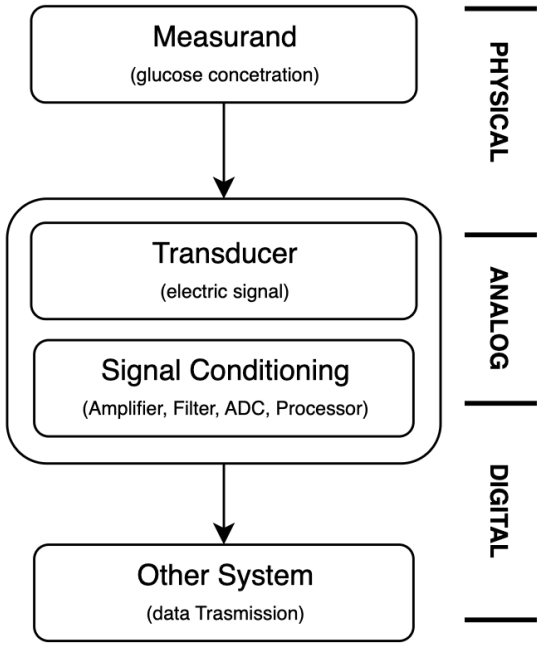


Figure 4.1. Measurement chain: from the physical to the digital domain

$$y(t) = g(u(t), t) + \eta(t) \tag{4.1}$$

In $g(u(t), t)$ and $\eta(t)$ it is possible to model the distorting factors and the noise. After a literature review, it is possible to list some of the factors to be considered when modelling the CGM sensor response.

1. *Temperature* can affect the transducer layer and the electronic layer. The kinematics of reactions that occur at the enzymatic level can be neglected because glucose oxidase denaturation occurs at 57.9°C and the body, in a living condition cannot reach that temperature [273]. Otherwise, the overheating of the device transmitted for conduction can lead to an error in reading due to the introduction of Johnson-

Nyquist noise. This noise can be simulated using white noise with an amplitude equal to the thermal noise power calculated using Nyquist's formula: $P = kTB$ where P is the noise power in Watt, k is Boltzmann's constant, T is the absolute temperature in Kelvin degrees, and B is the bandwidth of the system in Hertz.

2. *pH* also affects the glucose oxidate; its optimal activity is at 6.5, and when the pH is greater, the reaction speed decreases [274]. The signal acquired at a pH differing from the optimal one can be derived by multiplying the glucose concentration measured at the optimal pH by the ratio of enzymatic activities between the pH of interest and the optimal pH. Enzymatic activity can be deduced employing the Hill equation.

3. *Glucose rate of change*: The concentration of blood glucose and interstitial glucose, where CGM sensors take measurements, does not align instantly. There is a lag between the two. Sensors may consistently read higher or lower values depending on the rate at which glucose changes. Specifically, when the glucose rate is increasing, the sensors often report a lower value, and when it is decreasing, they report a higher value [275].

4. *Adhesion*: Movement or impact involving the subject can stress the needles or cause them to lose adhesion, leading to inaccurate signal readings. The sensor impact could be modelled as a pulse that interferes with only one sample belonging to the curve, whose value is altered by a variable multiplicative factor.

5. *Sensor Drift*: Sensor response changes over time due to multiple factors such as the biological body response that causes, electrode oxidation, and sensor degradation [276]. For these reasons, commercial sensors can measure glucose concentrations for a duration of $8-14$ days depending on device type [277]

The factors mentioned above are summarized in Table 4.1, where each

is associated with one or more potential causes such as the subject, environment, and sensor.

Table 4.1. Origin of type of error in CGM sensor

Type	Error	Subject	Environment	Sensor
Temperature	enzyme	-	x	-
Temperature	transducer	-	x	-
pH	enzyme	x	-	-
Drift	electrodes	-	-	x
Glucose rate of change	transducer	x	-	-
Adhesion	electrodes	-	x	-
Fabrication offset	all the parts	-	-	x

Methods

This section focuses on two important tools used in the work: (i) the UVA/PADOVA Type 1 Diabetes Simulator and (ii) the proposed Sensor Model.

Simulator The UVA/PADOVA Type 1 Diabetes Simulator is a sophisticated simulation software designed to facilitate the development and evaluation of treatment strategies for people with Type 1 Diabetes Mellitus (T1DM). Developed through collaborative efforts between the University of Virginia and the University of Padova, this simulator offers several notable features and functionalities. It replicates meal challenges and boasts a diverse population of 300 in-silico subjects, including adults, adolescents, and children, thereby enabling comprehensive research across different age groups. Importantly, the simulator has undergone rigorous validation against data from various T1DM experiments, successfully reproducing insulin correction distribution patterns. Continual improvement and evolution have been a hallmark of this simulator, culminating in the submission of a new version, S2013, to the FDA in 2013 [278].

Sensor Model The sensor model is mathematically represented by two main equations that describe the relationship between the sensor's internal state, the true Blood Glucose Level (BGL), and the sensor's output measurement. The model is described in [279] and [280]: it is composed of an equation (Eq- 4.2) that describes the state dynamic:

$$x_{k+1} = p_1 \cdot x_k + p_2 \cdot u_k \tag{4.2}$$

and a second equation (Eq- 4.3) that contains the output measurement formalization:

$$y_k = p_3 \cdot (x_k)^{p_4} \cdot \left\{ p_5 + \tanh\left(\frac{t_k}{3T} - 5\right) \cdot \mathrm{s}\left(\frac{t_k}{3T} - 5\right) \cdot \left[p_6 + \mathrm{s}\left(p_7 - \frac{t_k}{3T}\right)\right]\right\} \tag{4.3}$$

where $s()$ is the sigmoid function, T is the observation time expressed in seconds, and t_k is measured in seconds.

The dynamics of the sensor model are based on the zero-order-hold (ZOH) discretization of a continuous-time first-order stable transfer function from the true BG, u_k, to the internal state, x_k. This essentially means that the sensor's internal state is a discretized representation of the continuous changes in BG. There is a non-linear relationship between the sensor's output and its internal state, as well as the elapsed time. The sensor model also demonstrates the non-linear dependence of the output on the state and the non-linear drift of sensitivity over time. This means that the sensor's measurement is influenced by both its current state and the time that has passed since the start of measurements.

In the equation, some parameters appear, while in the reference literature, they are substituted by some specific value. The optimum parameters are evaluated using the Nelder-Mead method, which is a popular algorithm for multidimensional unconstrained optimisation without derivatives. The cost function minimized the quadratic error between the blood glucose concentration and the signal obtained with the model. A vector of ones is the starting point for the optimisation to find the minimum values. The simu-

lator was developed by customizing simglucose [281].

The simulator output data, blood glucose, and concentration u_k are given as input to the model equation, and each parameter was varied within a range of 5% from its nominal value. The results are presented as absolute errors. Data are obtained by simulating the blood glucose concentration of `adult#001` given the same bolus for 10 consecutive days. The scenario considered is reported in Table 4.2 and the optimal parameters are shown in Table 4.3.

Table 4.2. Meal size simulated for `adult#001`

Time [hr]	**Meal Size**
07:00	45
12:00	70
16:00	15
20:00	80
23:00	10

Table 4.3. Optimum parameters value

Parameter	p_1	p_2	p_1	p_2	p_1	p_2	p_1
Optimal value	-0.033	1.003	0.785	0.999	1.540	1.152	1.327

Application of distortions The values used for the application of distortion were obtained from literature or as a result of experimental tests. The temperature values analysed are: 20°C, 30°C, 40°C, 45°C. The pH values tested are: 4.5, 5.5, 6.5, 7. The glucose rate of change (GRC) tested are 1) : $(-0.5, -1)$, 2): $(-1, -1.5)$, 3): $(0.5, 1)$, 4): $(1, 1.5)$ [275]. The percentages of adhesion due to movement or impact involve a variation of the signal at the disturbance, affecting only one sample with a multiplicative factor of 1.5, 2, 2.5, and 3.

The sensor drift was modelled after an experimental test on a device

whose operating principle is based on amperometry, a technique that is present in various CGM devices. A square wave that oscillates between 0 mV and 0.360 mV, a frequency of 0.01Hz and a duty cycle of 50% was applied to an electrochemical cell. This cell is made of a commercial electrode Dropsense C110D with a 4 mm diameter working electrode in carbon, an auxiliary electrode in carbon and a reference electrode in silver [282]. The steady-state value was recorded for the various measurement cycles, and a distortion of the measurement was observed, which was modelled with a sixth-order polynomial. This effect was applied to the signal, and various levels of distortion were simulated by acting on the amplitude of the distortion, pre-multiplying it by a factor that specifically assumes values of 0.2, 0.5, 0.75, and 1.

Experimental Result

Model Characterization Taking into account Eq. 4.2, the parameters p_1 and p_2 can mimic factors that can be directly related to the transduction phase. With these parameters, it is possible to simulate the sensor calibration error, temperature and pH fluctuation, tissue differences, hydration, sensor age, physical pressure on the sensor, and sensor placement. In the parameter p_2 are mapped all the variables that directly influence the glucose concentration reading, such as pH, adhesion, sensor placement, and sensor manufacturing offset. Conversely, p_1 could capture the effects of aging. Specifically, as observed in Fig. 4.2, for p_1, a slight decrease in the parameter value corresponds to an increase in the absolute error. The effect induced by the variation of p_2 and p_3 is more intricate (Fig. 4.3). Throughout the day, the absolute error fluctuates, showing distinct patterns at different times. In the morning and afternoon, the absolute error tends to decrease when there is a reduction in the parameter variation. Conversely, in the evening, a decrease in parameter variation leads to an increase in the absolute error.

Eq. 4.2, with its parameters, allows one to model the effects of disturbances in the electrical domain. Specifically, during this phase, it is possi-

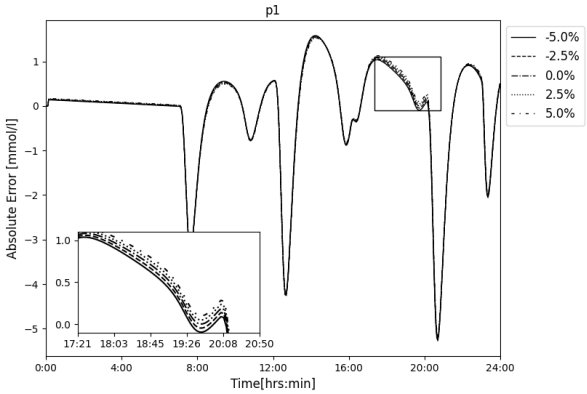


Figure 4.2. p1 parameter variation in the range of 5% from its nominal value

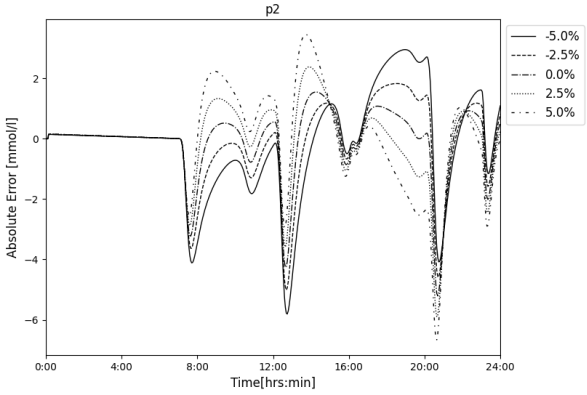


Figure 4.3. p2 parameter variation in the range of 5% from its nominal value

ble to consider all the disturbances that could interfere with the measured signal. Among these, thermal noise is certainly a prominent factor. The parameter p_4 appears as an exponent, it modulates the nonlinearity of the relationship between x_k and y_k. If p_4 is greater than 1, the function grows exponentially, this means that for increasing values of x_k , the increase of y_k becomes progressively faster. When $p_4 < 0$ and x_k increases, y_k approaches zero. The greater the value of x_k , the smaller y_k becomes, although it never

actually reaches zero (Fig. 4.4).

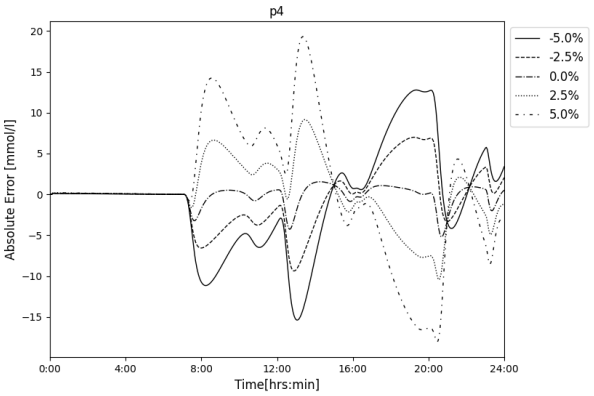


Figure 4.4. p4 parameter variation in the range of 5% from its nominal value

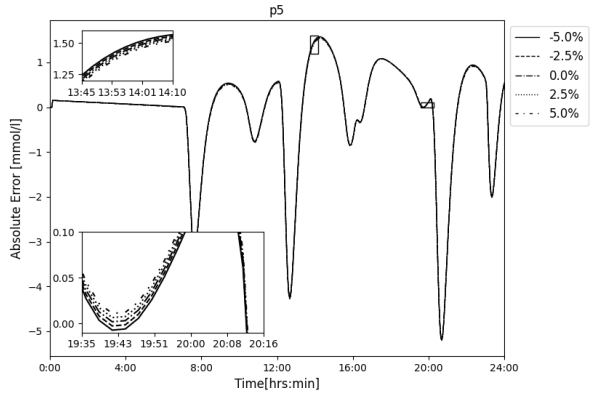


Figure 4.5. p5 parameter variation in the range of 5% from its nominal value

The parameters p_5 and p_6 do not influence the absolute error (Fig. 4.5) whereas p_7 exhibits behaviour analogueous to p_2.

Model characterization for the case study In this section, for each interfering substance, its response has been modelled, and the range of parameter variations within the model has been examined to establish a cor-

responding dependency between the parameter and the case study. This approach ensures that the influence of each interferent on the system's accuracy is not only identified but also quantitatively assessed.

Table 4.4. Model Parameters for the Considered Interferents

Interference	Value	p_1	p_2	p_3	p_4	p_5	p_6	p_7	Error
Adh	1.5	-0.034	1.002	0.790	1.000	1.549	1.164	1.334	-0.042±3.134
Adh	2	-0.034	1.005	0.789	1.000	1.558	1.161	1.335	0.217±3.307
Adh	2.5	-0.033	1.007	0.795	0.998	1.554	1.180	1.333	0.117±6.485
Adh	3	-0.034	1.004	0.791	1.000	1.542	1.160	1.329	0.129±6.008
pH	4.5	-0.031	1.152	0.793	1.062	1.499	1.131	1.267	-0.024±0.278
pH	5.5	-0.033	1.013	0.790	1.001	1.542	1.167	1.337	-0.154±0.508
pH	6.5	-0.034	1.006	0.792	0.999	1.553	1.162	1.332	0.31±1.206
pH	7	-0.034	1.006	0.790	1.000	1.554	1.165	1.331	0.031±0.857
T [°C]	20	-0.034	1.006	0.792	0.999	1.553	1.162	1.332	0.31±1.206
T [°C]	30	-0.034	1.006	0.792	0.999	1.553	1.162	1.332	0.31±1.206
T [°C]	40	-0.034	1.006	0.792	0.999	1.553	1.162	1.332	0.31±1.206
T [°C]	45	-0.034	1.006	0.792	0.999	1.553	1.162	1.332	0.31±1.206
GRC	0	-0.033	1.009	0.791	0.999	1.568	1.169	1.324	-0.064±2.425
GRC	1	-0.033	1.006	0.789	1.000	1.565	1.168	1.327	-0.037±2.299
GRC	2	-0.034	1.005	0.789	0.999	1.550	1.161	1.336	-0.041±4.184
GRC	3	-0.034	1.006	0.785	0.999	1.554	1.163	1.342	-0.158±2.339
Drift	0.2	-0.034	1.005	0.791	0.999	1.553	1.162	1.332	0.219±1.213
Drift	0.5	-0.034	1.005	0.789	1.000	1.543	1.166	1.334	0.228±1.222
Drift	0.75	-0.034	1.007	0.787	1.000	1.548	1.165	1.332	0.081±1.26
Drift	1	-0.034	1.004	0.788	1.000	1.557	1.164	1.334	0.18±1.223

In Table 4.4 for each row is reported the interference taken into account and its value. All values of interferents given in the table, excluding temperature, are expressed as multiplicative factors and therefore dimensionless. From columns p1 to p7 are reported the value of the model that better fitted the signal affected by that interference. The last column shows the average and the standard deviation of the error obtained as the difference between the signal and the model curve.

Discussion and Conclusion

The analysis demonstrates that the parameter variations observed fall within the expected range of the analysed parameters, with several noteworthy patterns emerging from the model's performance. A particularly significant finding is the consistent behaviour of parameter p4, which maintains stability across all case studies. The model's effectiveness in representing adhesion processes is evidenced by the standard deviation error between the observed and modelled signals, with this divergence being most pronounced in parameters p2 and p4. The model reveals important relationships between environmental factors and parameter behaviour. pH levels demonstrate a clear influence, causing a reduction in parameters p2 and p4 as pH increases. Conversely, temperature variations within the studied range show minimal impact on these parameters, suggesting that temperature plays a negligible role under the tested conditions. The model also identifies an inverse correlation between glucose rate changes and parameters p2, p5, and p6, with these values decreasing as glucose rates increase. In scenarios involving sensor drift, all parameters except p1, p4, and p7 show variations, highlighting specific parameter sensitivity to drift conditions. These findings have important implications for continuous glucose monitoring (CGM) systems. The model equation's parameters effectively capture and represent the impact of interferents on blood glucose concentration measurements. The observation that certain parameters display similar values across different effects provides valuable insights for sensor calibration and compensation strategies. This understanding is crucial for maintaining CGM sensor accuracy and reliability across varying operational conditions. The identification and calibration of these parameters emerge as critical factors in ensuring robust CGM sensor performance. By accounting for these parameter behaviours, it becomes possible to maintain measurement accuracy despite the dynamic and potentially disruptive external conditions that sensors may encounter during operation. This understanding provides a foundation for developing more resilient and accurate glucose monitoring systems that can adapt to varying physiological and

environmental conditions while maintaining measurement integrity.

4.3.3 A Stress Generation Model for Tiny ML Drift Compensation

Background and Research Questions

Several studies have focused on modeling the inherent errors associated with GCM sensors, shedding light on critical aspects that impact their accuracy and reliability.

Krouwer and Cembrowski emphasized the need for standards and statistics to describe the performance of the blood glucose monitor [266] in a comprehensive way. This laid the groundwork for a holistic approach to the evaluation of CGM sensors, considering analytical errors and addressing medical errors that could potentially harm patients. Facchinetti *et al.* delved into modeling the glucose sensor error [267]. Their work highlighted the challenges faced by CGM sensors, citing distortions due to diffusion processes, time-varying systematic under/overestimations from calibrations and sensor drifts, and the presence of measurement noise. In another study, Facchinetti identified and evaluated error models for the G4 Platinum (G4P) and advanced G4 for artificial pancreas studies [268]. In their research, the authors highlighted the technological progress, with the G4P surpassing its forerunner, the SEVEN Plus, in performance, and the G4AP exhibiting enhanced reliability due to advanced data processing algorithms. Vettoretti *et al.* broadened the examination to include self-monitoring blood glucose (SMBG) measurements, introducing an innovative method for developing more accurate models of SMBG error probability density functions (PDFs). Their study segmented the blood glucose spectrum into zones, each defined by a consistent standard deviation. This novel strategy overcame the shortcomings of conventional Gaussian models and offered a more precise depiction of the experimental data.

With the diffusion of factory-calibrated CGM sensors, Vettoretti developed a model that dissects the error into BG-to-IG kinetics, calibration er-

ror, and measurement noise [270]. This model extended the applicability to the entire sensor lifetime, a significant advancement considering the 10-day duration of these sensors. In long-term glucose forecasting, Liu *et al.* proposed an algorithm based on physiological models and the deconvolution of CGM signals [271]. Their research tackled the difficult task of making accurate long-term forecasts, an essential component for applications like precision insulin dosing and artificial pancreas systems. Facchinetti et al. (2013) utilized real data from multiple simultaneous CGM recordings of Dexcom SEVEN Plus sensors, alongside frequent BG references, to propose a model describing CGM sensor errors without distinguishing between physiological and technological errors. They reported a Mean Absolute Relative Difference (MARD) with an average global MARD of 14.2%, including contributions from the BG-to-IG diffusion process (3.5%), calibration errors (12.8%), and measurement noise (5.6%) [267]. Drecogna et al. (2021) used real data from 167 adults with the Dexcom G6 sensor to model data gaps in CGM sensor data due to temporary sensor errors or disconnections, employing a two-state Markov model for parameter estimation [283]. Talukder et al. (2022) utilized datasets from live rats and FDA-approved virtual diabetic patient models to develop a Bayesian inference-based nonlinear, non-causal dynamic calibration method for sensors with nonlinear, time-drifting characteristics, achieving estimation errors within 9.83% of true BG values [280].

This work aims to develop a novel modelling approach for predicting sensor calibration loss and implementing automated compensation strategies. We propose a new database architecture that captures the complex relationships between environmental factors, physical wear, and material degradation to enable proactive calibration adjustments. Unlike previous studies that rely on clinical data from commercial Continuous Glucose Monitoring (CGM) sensors, our approach leverages simulated data from a public simulator enhanced with specific interference effects. This allows us to systematically evaluate sensor behaviour across a comprehensive range of operating conditions, with particular focus on modelling a sensor family that

accurately reflects the error distribution patterns observed in commercial devices. Our dataset incorporates multiple physiological and technological effects to ensure realistic sensor response simulation. Our model integrates commercial device characteristics with documented interference patterns from literature to ensure real-world applicability. Additionally, we demonstrate the practical value of our dataset by implementing a non-neural machine learning algorithm optimised for low-power microcontrollers (MCUs), validating the feasibility of deploying ML-based compensation strategies in resource-constrained environments.

Materials and Methods

Model Description In this study, the model processes input data derived from 500 simulations of 10 adult individuals. For each subject, an interval of 15 days glucose response was generated, in accordance with the lifetime of CGM sensors, which varies between 8 and 14 days.

The generation of a daily meal plan mimics the dietary patterns of an individual and it is called a scenario. It defines the probabilities for different meals throughout the day (breakfast, two snacks, lunch, dinner, and a third snack), along with their usual time ranges and nutritional amounts. For each meal, a truncated normal distribution is used to determine the meal time, ensuring that it falls within realistic bounds, while the amount of meal is determined based on a normal distribution.

Upon the conclusion of each day, the scenario undergoes a reset to initiate a new cycle, this approach enables the generation of several meal distributions across different days, both in terms of quantity and timing. Based on the scenario and the characteristics of the patient, the simulator gives the response in terms of blood glucose concentration over time, $BG(t)$, generating this value every 3 minutes.

The sensor response combines several contributions that mimic the real effect of the measurement process on the analyte. It can be described by the following equation:

$$CGM(t) = IG(t) + \eta(BG(t)) + \xi(t) + \epsilon(t) \tag{4.4}$$

where $IG(t)$ captures the blood glucose-to-interstitial glucose (BG-to-IG) kinetics, $\eta(a)$ shows the measurement sensor error based on the data reported from commercial sensors, $\xi(t)$ represents the white noise affecting the measure and $\epsilon(t)$ is the sensor drift over the time. All these effects are represented in a block diagram in Figure 4.6.

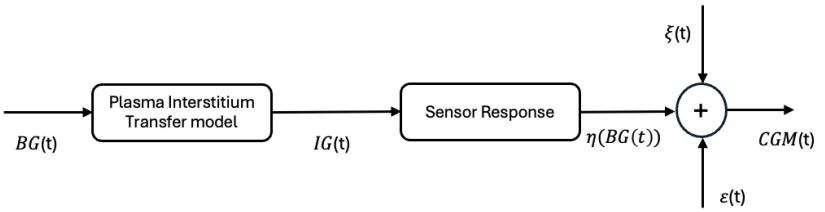


Figure 4.6. Graphical representation of the sensor response: $BG(t)$ - blood glucose concentration , $IG(t)$ interstitial glucose concentration, $\eta(a)$ - measurement sensor error; $\xi(t)$ - white noise and $\epsilon(t)$ - sensor drift over the time.

Going deeper into the specific elements of the equation:

- $IG(t)$: this aspect was deeply studied in literature. It indicates the diffusion of glucose from the blood to the measurement site of CGM: the interstitium. This diffusion process causes an attenuation of the amplitude and phase delay of the $IG(T)$ compared to the $BG(t)$ profile. The time constant τ that describes this process has a variability within and between subjects and ranges from 6 to 15 minutes [284]. In this study, based on the equation reported in [285], it is calculated considering also the previous value of estimated $IG(t)$:

$$IG(t) = IG(t-1) \cdot \exp\left(-\frac{t}{\tau}\right) + BG(t-1) \cdot \left(1 - \exp\left(-\frac{t}{\tau}\right)\right) \tag{4.5}$$

- $\eta(a)$: the measurement sensor error is introduced into the model to characterize the commercial sensor response. Although the commonly used simulator employs global metrics to evaluate the behavior of a specific sensor, this model has set its goal to achieve a response that more accurately mimics the real signals obtained from the sensors. As the technical user guides report, when the sensors are used, they show an error compared to the reference measure obtained from a gold standard blood glucose. Data obtained in a clinical study were compared with the response of the Yellow Springs Instrument 2300 STAT PlusTM glucose analyse (YSI). In this way, from the sensor datasheet it was possible to obtain the concurrency of the measurement and measurement error on a group of adult subjects [286]:

Table 4.5. Concurrence of G7 sensor readings and YSI values by YSI glucose range obtained from 308 adults expressed in % [286].

CGM	YSI value range [mg/dL]										mg/dL>mg/dL
	<40	40-60	61-80	81-120	121-160	161-200	201-250	251-300	301-350	351-400	>400
<40	61.54	4.71	1.53	0.04	0.06	0.00	0.00	0.00	0.00	0.00	0.00
40-60	34.62	63.56	14.85	0.99	0.04	0.00	0.00	0.00	0.00	0.00	0.00
61-80	3.85	29.45	65.02	9.44	0.21	0.10	0.00	0.00	0.00	0.00	0.00
81-120	0.00	2.20	18.48	77.51	12.57	0.71	0.18	0.00	0.00	0.00	0.00
121-160	0.00	0.09	0.03	11.90	74.01	15.15	1.46	0.13	0.00	0.00	0.00
161-200	0.00	0.00	0.09	0.13	12.95	69.46	16.85	1.42	0.09	0.00	0.00
201-250	0.00	0.00	0.00	0.00	0.17	14.48	67.77	16.82	1.11	0.34	1.69
251-300	0.00	0.00	0.00	0.00	0.00	0.10	13.57	67.27	25.91	4.37	0.00
301-350	0.00	0.00	0.00	0.00	0.00	0.00	0.18	14.13	61.01	33.90	3.37
351-400	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	11.57	53.29	26.97
>400	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	8.11	67.98

To simulate the sensor response, for each YSI interval of values reported in the Table 4.5 as columns, based on the probabilities reported in each row of the same column, the CGM upper limit in the interval is calculated as

$$\mathrm{CGM}_i \sim f_{\mathrm{CGM}_i}(.) \quad \forall i \in \{1, 2, \ldots, 11\}$$

where

$$f_{\mathrm{CGM}_i}(.) = \sum_{j=1}^{11} P_{i,j} \times \mathscr{U}\left(\mathrm{mCGM}_j, \mathrm{MCGM}_j\right)$$

In the above equation, the values mCGM_j and MCGM_j represent the minimum and maximum limits of the interval corresponding to row j.

This operation is repeated for all intervals, ensuring an increasing response in the extraction process. If $\{(CGM_i, YSI_i)\} \quad \forall i \in \{1, 2, \ldots, 11\}$ is the set of points obtained above, the values between them are computed based on linear interpolation as:

$$CGM(x) = CGM_i + \frac{(CGM_{i+1} - CGM_i)}{(YSI_{i+1} - YSI_i)}(x - CGM_i) \quad \forall x \in [YSI_i, YSI_{i+1}]$$

In Figure 4.7 there is a representation of how the sensor response is obtained. The dotted lines represent the thresholds at which the pairs of values are determined (CGM_i, YSI_i) while the red line gives the complete sensor response resulting by the linear interpolation.

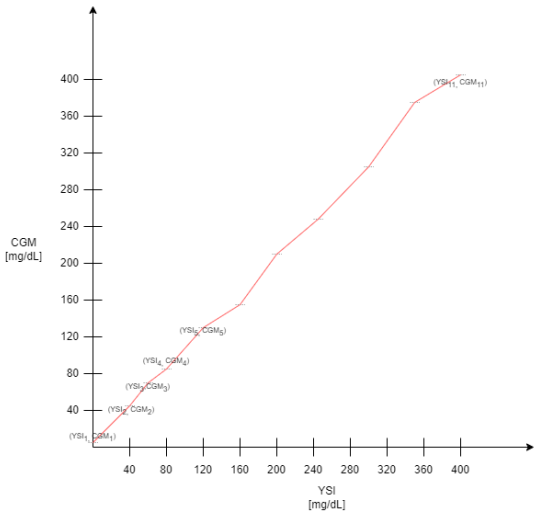


Figure 4.7. Sensor response, the dotted lines represent the extracted values, in red the linear interpolation between these points.

To better describe this process, the algorithm is described with pseudocode (Algorithm 1). In the algorithm description *bins* represents the number of intervals that can be defined in the YSI values, that in this specific case is equal to 11.

Algorithm 1 Generate CGM Data Points with Adjusted Concentrations

Require: $bins$, YSI, mCGM, MCGM, P
Ensure: CGM, interpolator
Initialize CGM as an empty array
Initialize $previousIndex$ as -1
for $i = 1$ to $bins$ **do**
    $currentIndex \leftarrow$ Select a random index from 1 to 11 using column i probabilities
    if $currentIndex \leq previousIndex$ **then**
        $currentIndex \leftarrow previousIndex$
    end if
    $currentRange \leftarrow [mCGM_{currentIndex}, MCGM_{currentIndex}]$
    $previousIndex \leftarrow currentIndex$
    $CGM \leftarrow$ Generate random value within $mCGM_{currentIndex}, MCGM_{currentIndex}$
    Append CGM_i to CGM
end for
$interpolator \leftarrow$ Create linear interpolator from YSI and CGM
return $CGM, interpolator$

- $\xi(t)$: the noise that affects the measure is defined as white noise with an amplitude included in ±5% of the measure.

- $\epsilon(t)$ sensor response changes over time due to multiple factors, such as the biological body response that causes electrode oxidation and sensor degradation [287]. For these reasons, commercial sensors can measure glucose concentrations for a duration of 8-14 days depending on the type of device [277]. In this paper, the drift is modeled as a linear effect in which the slope of the line is obtained based on the range of values on the first day of observation, to simulate the effect reported in [288].

Model Evaluation After producing the dataset as described in the previous section, the model is evaluated in terms of the distribution of obtained error and problem complexity. Two classic machine learning (ML) algorithms are presented, to understand the complexity of the problem: the Random Forest Regressor (RFR) and the Support Vector Regressor (SVR).

The RFR is based on multiple decision trees to predict a continuous outcome. It operates by constructing a multitude of decision trees at training time whilst outputting the average prediction of the individual trees to form a more accurate and robust prediction. It is widely used for various regression tasks due to its simplicity, ease of use, and ability to capture non-linear relationships between features and the target variable.

The SVR is a type of Support Vector Machine (SVM) that is used for regression tasks. It works by finding the hyperplane that best fits the data in a high-dimensional space, trying to minimize the error within a certain margin. It is capable of capturing complex, non-linear relationships by employing different kernel functions (linear, polynomial, radial basis function, etc.) to map input features into higher-dimensional spaces.

The RFR is evaluated considering different maximum values of parameters, referred to within the paper as max depth, in the range of 1-9, while SVR is evaluated using the linear kernel. For both the ML algorithms the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the R^2 are evaluated and compared.

All the analyses presented in this paper are performed considering the dataset that is reshaped into one-dimension array. Results were obtained on Google Colab using as runtime Python 3 on a A100 GPU accelerated hardware.

Model on Specific Sensor To evaluate the specificity of each sensor response, this section wants to analyse how the goodness of the sensor error response can be predicted given a single sensor output. This analysis was carried out using the RFR optimised in terms of MSE to get the number of estimators and parameters. After optimising the model, it is applied to

each record, which represents a sensor's data collected over a 15-day period. Each derived dataset, indeed, is composed of a 500x2 matrix where the first column represents the time and the second column the sensor error, while the ground truth is represented by the CGM value given by the simulator. Specifically, the initial 80% of the data trace from the specific sensor is employed to train the model. The remaining 20% of the data trace, which corresponds to the latter part of the recording period, is then used to test the model. This segment is particularly critical as it includes data where prediction errors can have more significant consequences, possibly due to the accumulation of small variances over time or abrupt changes in sensor behavior. This methodological approach of dividing the data is systematically applied across all 500 sensors involved in the study. Training and testing the model on these segmented portions of data from each sensor ensures that the model is both robust and capable of handling real-world variabilities in sensor outputs. Upon completion of the training and testing cycles, the MAE and MSE are calculated for each sensor's predictions. The results are then aggregated to derive mean values and relative standard deviations for these metrics across all sensors. This statistical analysis provides a clear overview of the model's overall accuracy and reliability in predicting sensor responses, highlighting its strengths and areas for potential improvement.

Results

The model is used to generate a dataset of 500 sensor responses for a duration of 15 days. The dataset has a final dimension of 500x7200. In Figure 4.8 there are a bunch of sensor responses in the range of 0 - 500 mg/dL overlapped over an ideal sensor response (dashed red line). Figure 4.9 shows a compact representation of all the 500 responses where the blue line shows the average sensor response, and the area is the 1^{st} and 3^{rd} quartile. It can be noticed that for low and high values of blood glucose concentration, the majority of measures cover a range wider than the region of interest (70 mg/dL - 180 mg/dL).

An example of a signal resulting from the model generated over the 15

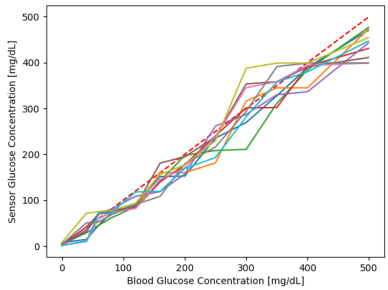


Figure 4.8. 10 sensor responses, the dashed line is the bisector that represents the ideal sensor response

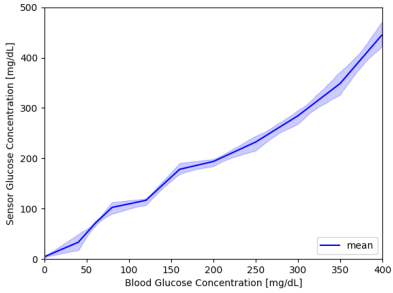


Figure 4.9. 500 sensors response, the blue line shows the average sensor response, while the area covers the 1^{st} and 3^{rd} quartile.

days is shown in Figure 4.10, while a representation in terms of mean, 1^{st} and 3^{rd} quartile is reported in Figure 4.11. The absolute error grows over time reflecting the degradation of the sensor.

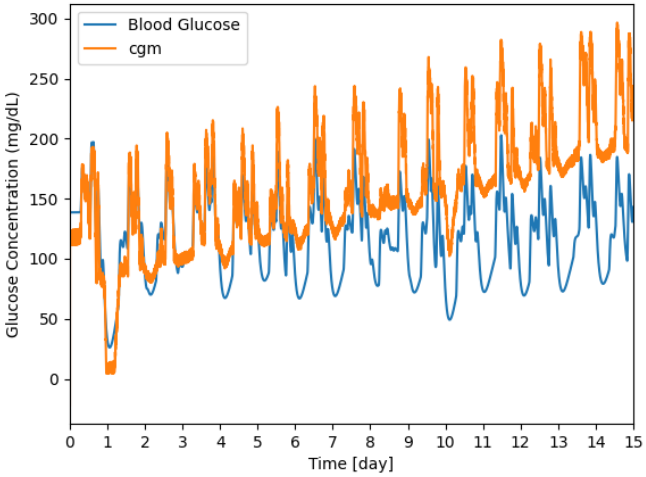


Figure 4.10. Example of a signal generated by the model for 15 days, in orange the CGM sensor response and in blue the reference signal

The absolute error evaluated as the average of every sensor response is

expressed as Cumulative Distribution of Frequency (CDF), Figure 4.12. This graph shows that the error seems to be greater than zero. The distribution has a mean value of 40.79 mg/dL, with the 25^{th} percentile at 21.02 mg/dL and the 75^{th} percentile at 58.46 mg/dL.

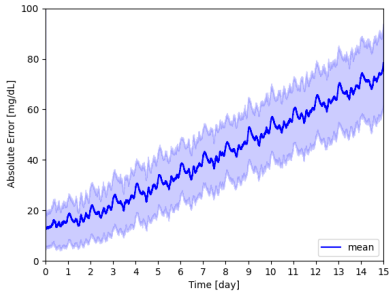


Figure 4.11. Absolute glucose concentration error, in blue the mean value over time, while the area represents the measures that are within the 25^{th} and 75^{th} percentile.

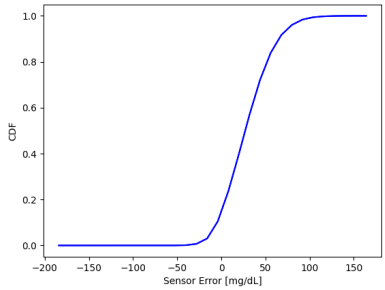


Figure 4.12. Cumulative distribution of sensors error. Mean = 40.79 mg/dL, the 25^{th} percentile is at 21.02 mg/dL and the 75^{th} percentile is at 58.46 mg/dL.

Model Evaluation The dataset is reshaped in a one-dimensional array and it is split in train and test sets. The RFR model is trained using a max depth of 3 and 100 trees, while the SVR is trained with a linear kernel. Below the obtained results:

Table 4.6. Root Mean Square Error ($RMSE$), Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) for RFR and SVR

Model	RMSE [mg/dL]	MSE [(mg/dL)^2]	MAE [mg/dL]	R^2 [(mg/dL)^2]
RFR	20.55	422.24	16.13	0.42
SVR	20.90	436.89	16.22	0.41

The results obtained with RFR are slightly better than SVR, although still not satisfactory for the pre-set task.

Regarding Table 4.6, the presented results may seem unusual, but this analysis was conducted to grasp the complexity of the problem. Deriving the relative value of blood glucose concentration from the sensor model using traditional machine learning algorithms is challenging. This difficulty highlights the complexity of the dataset and underscores the need for more advanced artificial intelligence algorithms.

Model on Specific Sensor The last analysis made it possible to evaluate the performance of the trained models on the responses of individual sensors. The results obtained in terms of MSE is $223.93 \pm 234.11[(mg/dL)^2]$ and MAE $11.01 \pm 5.12[mg/dL]$. The latter result demonstrates the strong variability of traces within the dataset, whereby some specific models on some sensors perform very well, while others perform very poorly, as reported by the high standard deviation on the MSE metric.

Deployability on MCU

Considering that the CGM task requires an implementation of algorithms on a microcontroller (MCU) or a sensor itself, the definition of the mandatory requirements for the applicability of a model on such devices must be calculated. After choosing the max depth that results in less than 1% decrease in error, the portability of this model on the MCU is evaluated. From the graph shown in Figure 4.13, it's noticeable that by increasing max depth, there are no great improvements in terms of RMSE. This suggests that due to the complexity of the dataset, probably more complex algorithms of regression are needed.

To assess the model's portability on MCU, the STM32Cube.AI Developer Cloud tool [289] is utilized. This tool, freely available on the website, enables a machine learning developer to upload a pre-trained machine learning workload. Next, it allows to automatically profile its computational complexity and memory requirement. Then it enables the automated ANSI C code conversion of the imported model. The C code is transparently integrated into a built/in application which is then compiled and installed on

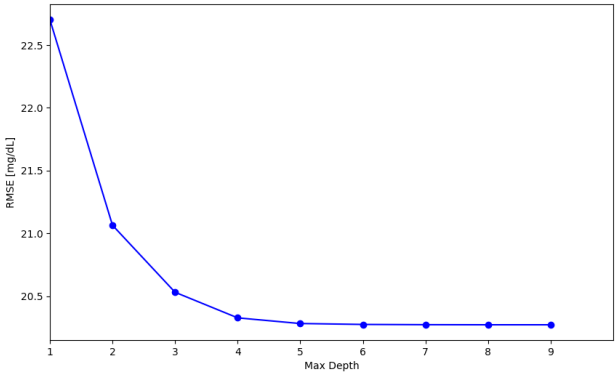


Figure 4.13. RMSE evaluation for RF models with a variation of the model max depth parameter

the MCU chip. Finally, the installed firmware will be executed and detailed profiling (including the execution time) made available to the user. With this solution it is possible to test the developed algorithm on the MCU chip. This tool facilitated the estimation of the selected model's requirements, including the number of Multiply-Accumulate Operations (MACCs), as well as the needed flash memory, RAM space, and the inference time. The board U5855I is selected, due to its low power consumption (Arm Cortex-M33, 160MHz, 768KB of RAM for AI). The number of MACCs is 800, the flash needed is 24 KiB, including 15.43 KiB for network weights and about 8 KiB for libraries, and the total RAM is about 2KiB. Benchmark on the selected board determined an inference time of 0.1920 ms.

The complete result is shown in Table 4.7:

Table 4.7. Details of algorithm implementation using the optimisation of STM32Cube.AI Developer Cloud tool

Software version	optimisation	Allocation Inputs	Allocate Output	MACC	Flash size	RAM size
9.0.0	RAM	TRUE	TRUE	800	Total: 24 KiB Weights: 15.43 KiB Library: 8 KiB	Total: 2 KiB Activation: 20 B Library: 2 KiB In/Out: 0B/0B

Discussion and Conclusion

Accurate monitoring of blood glucose is crucial for diabetes management and can significantly influence treatment decisions and patient well-being. A dataset that accurately mimics the behaviour of an array of sensor types becomes an essential tool for progress. Such a dataset aids in developing robust machine-learning algorithms that can handle intrinsic variability and potential sensor degradation. Traditional machine learning methods, such as RFR and SVR show promise but also exhibit limitations, particularly regarding systematic and random errors even when these are trained on data from individual sensors reveal that the outcome is heavily dependent on the quality of the initial signal. This is probably due to the complexity of the task. One of the goals of this paper is to characterize the complexity of the generated dataset. While the presented analysis aligns with the model analyses presented in [267], which are based on real-world data, it demonstrates greater complexity compared to [270]. Unlike the existing literature, which primarily relies on real data, the presented approach is novel in that it utilizes simulated data to perform stress tests.

4.3.4 optimising Glucose Sensor Calibration with Lightweight Neural Networks: A Comparative Study

Background and Research Questions

Recent research endeavours have leveraged Machine Learning (ML) techniques to enhance CGM accuracy. For instance, a study employed a stacked long short-term memory (LSTM)-based deep recurrent neural network model to predict blood glucose levels [290]. Another study utilized an LSTM-based neural network to predict glucose levels for up to 60 minutes using continuous glucose measurements [291]. Additionally, a study titled "Exploring non-invasive features for continuous glucose monitoring" conducted at the University of Memphis summarized various minimally invasive and non-invasive sensors for accurate blood glucose measurement using different

ML models such as Linear Regression, Support Vector Regression (SVR), k-nearest Regression, Decision Tree Regression (DT), Bagging Trees Regressor, Random Forest Regressor (RFR), Gaussian Process Regression, and Multilayer Perceptron (MLP) [292].

However, the accuracy of non-invasive blood glucose monitors (NIBGMs) remains a challenge, particularly in measuring low glucose concentrations, due to within- and between-patient variations. Thus, it is imperative to scientifically partition the entire blood measurement range into smaller clusters or groups and train ML models for each cluster separately to accurately predict a non-invasive value.

This study aims to design and test a model capable of predicting the time error in the blood glucose level reading of a CGM sensor with high accuracy and minimal computational burden. The ultimate goal is to develop a model that could potentially be embedded directly into the sensor itself in the future, leveraging Tiny Machine Learning. A comprehensive investigation into the performance of various neural network architectures on the provided database was performed to achieve this objective. Subsequently, two models based on their performance metrics were selected and the variation of their error in glucose level prediction over the frequency of sensor calibration was evaluated. This analysis provides valuable insights into the efficacy of different calibration strategies in improving prediction accuracy.

Blood Glucose Level Simulation and Database Creation

The first step in creating the model involved selecting the database for blood glucose levels. This study used the dataset presented in the previous section obtained through simulation methods rather than one composed of real patient data [265].

Once the database was established, the predictive dataset was constructed as follows. Two key features were selected: time, with a granularity of 3 minutes, and model accuracy, quantified as the deviation between the actual and predicted values of the initial element within each temporal series for every sample. The ground truth comprised the temporal deviation be-

tween actual and predicted values.

Model Comparison for Time Error Estimation

Once the database was created, four types of models were trained and tested to predict the error in the sensor. All the models were designed for sequential data handling and capable of capturing long-term dependencies within sequences and were implemented using TensorFlow's framework for Deep Learning, Keras [293]. Specifically, the models implemented were:

1. Long Short-Term Memory (LSTM) [294]: it is a type of recurrent neural network (RNN) designed to handle sequence problems and store long-term information. Compared to traditional RNNs, LSTMs can address the vanishing or exploding gradient problem during training on long sequences. This is achieved through "memory cells" within LSTMs, which control the flow of information through the network and enable the retention and updating of long-term information.

2. Gated Recurrent Unit (GRU) [295]: it is another variant of recurrent neural network designed to address the vanishing gradient problem in RNNs. Compared to LSTMs, GRUs have a simpler structure with fewer parameters but are still capable of capturing long-term dependencies in sequences. GRUs use an "update gate" mechanism to control the flow of information within the network.

3. Legendre Memory Unit (LMU)[296]: it is a type of neural network that utilizes Legendre basis functions to capture and store long-term information. Compared to traditional architectures like LSTM and GRU, LMU offers advantages in terms of computational efficiency and long-term information storage capacity. LMUs are particularly suitable for processing temporal sequences and have been successfully used in a variety of applications, such as speech recognition and temporal forecasting.

4. Temporal Convolutional Network (TCN)[297]: it is a type of convolutional neural network specifically designed for processing temporal sequences. Unlike recurrent architectures like LSTM and GRU, TCN utilizes convolutional operations along the temporal axis to capture temporal dependencies in the data. TCNs have shown promising results in various sequence modeling tasks, offering advantages such as parallelism, efficiency, and ease of implementation.

The exploration embarked on a comprehensive analysis of diverse model architectures to scrutinize their configurations and performance across various types. These architectures were meticulously tailored to leverage the unique strengths of each model while addressing specific challenges within the dataset.

Each model was trained on time and initial error of device (simulating one-time pricking) to predict the time error and underwent configuration, employing MSE loss, Adam optimiser, and evaluation metrics encompassing both Mean Absolute Error (MAE) and Mean Squared Error (MSE). The training regimen spanned 150 epochs, ensuring thorough learning and optimisation. Subsequently, the models were trained on the dataset, with continuous assessment on validation data. Upon completion, the best-performing weights were evaluated on a separate test dataset, yielding comprehensive performance metrics as summarized in Table 4.8.

It is noteworthy that the TCN architecture consistently achieved lower MAE and MSE values across several instances, despite requiring a considerable number of parameters and occupying significant space. However, its hardware-friendly nature makes it a viable candidate for on-device implementation, potentially on a glucose sensor. Similarly, the LMU-based model (model 1) demonstrated optimal results with a notably lower parameter count, rendering it an attractive choice for integration into glucose sensors, minimizing energy consumption and spatial requirements.

Models such as model 11 and model 1 emerge as pivotal candidates for sensor integration due to their efficient performance characteristics. Other models, while demonstrating varying degrees of efficacy, may necessitate

Table 4.8. Summary of model architectures and performance metrics. Model 11 demonstrates superior performance in terms of MAE and MSE, while Model 1 offers a balance with lower parameter count, suitable for embedded systems. (Part 1)

Model Number	Model Category	Number of Parameters	Batch Size	MSE (mg^2/dL^2)	MAE (mg/dL)
model_1	LMU	296	65536	1044.44	25.34
model_2	LMU	94210	65536	1215.00	28.43
model_3	GRU	4421	65536	1326.46	29.22
model_4	GRU	5684	8192	1502.99	31.09
model_5	GRU	4400	32768	1796.95	34.43
model_6	GRU	4484	16384	2661.71	42.89
model_7	GRU	304718	4095	155748.55	106.47
model_8	LSTM	3429	65536	1170.32	27.16
model_9	LSTM	302734	4095	2584.02	41.13
model_10	LSTM	288386	1024	3592.58	52.16
model_11	TCN	1770	65536	974.54	24.44
model_12	TCN	819974	4096	1029.79	25.08
model_13	TCN	1242	65536	985.23	25.18
model_14	TCN	646014	4096	1010.20	25.41
model_15	TCN	39250	32768	1044.16	25.44
model_16	TCN	307610	4096	1052.49	25.72
model_17	TCN	98470	4096	1070.20	26.18
model_18	TCN	560738	4096	1119.02	26.58
model_19	TCN	48049	65536	1139.71	26.89
model_20	TCN	1033257	4096	1157.78	27.13
model_21	TCN	1598	65536	1195.75	27.37

Table 4.9. Summary of model architectures and performance metrics. Model 11 demonstrates superior performance in terms of MAE and MSE, while Model 1 offers a balance with lower parameter count, suitable for embedded systems. (Part 2)

Model Number	Model Category	Number of Parameters	Batch Size	MSE (mg^2/dL^2)	MAE (mg/dL)
model_22	TCN	141716	4096	1203.02	27.52
model_23	TCN	77496	32768	1228.10	27.61
model_24	TCN	56030	32768	1237.88	27.70
model_25	TCN	18014	65536	1307.52	29.28
model_26	TCN	12906	65536	1480.36	30.83
model_27	TCN	1586	65536	1539.88	31.26
model_28	TCN	82066	32768	1501.74	31.28
model_29	TCN	770	65536	1555.90	32.10
model_30	TCN	18150	65536	1609.55	32.10
model_31	TCN	294934	4096	1701.29	33.77
model_32	TCN	680730	4096	2184.94	36.48
model_33	TCN	131678	4096	2139.18	38.08
model_34	TCN	1770	65536	2323.57	39.32
model_35	TCN	8490	65536	2380.07	40.38
model_36	TCN	317050	4096	2494.26	40.89
model_37	TCN	3234	65536	2881.17	42.21
model_38	TCN	54	65536	2748.48	44.91
model_39	TCN	1024	65536	3710.99	51.09
model_40	TCN	76448	32768	4026.26	51.15
model_41	TCN	10698	65536	5863.49	54.12
model_42	TCN	3108	65536	6639.39	62.59

further optimisation or consideration based on specific application requirements. The summarized performance metrics of the two most promising models can be found in Table 4.10.

Table 4.10. Models that obtained the best Performance Metrics

Model Number	Model Category	Number of Parameters	MSE (mg/dL)	MAE (mg^2/dL^2)
model_1	LMU	296	1,044.44	25.34
model_11	TCN	1,770	974.54	24.44

Evaluation of Performance Based on Sensor Calibration Frequency

In the previous section, our assessment of the models focused on their accuracy computed on the initial sample of the time series, denoting the first instance of device usage. This section delves into an examination of model performance, centring on the frequency of sensor calibration. For this investigation, two models are singled out: model 11, distinguished by its superior performance in terms of minimal MAE and MSE in the prior analysis, leveraging a TCN architecture with 1770 trainable parameters; and model 1, an LMU-based architecture, which, despite marginally higher MAE and MSE values, operates with substantially fewer parameters, totaling 296.

Each of the models underwent training for 150 epochs, mirroring the approach adopted in the previous phase. Utilizing MSE loss, Adam optimiser, and evaluating both MSE and MAE, the models were trained across a spectrum of 1 to 10 calibrations within the 15-day time series, ensuring a homogeneous distribution of calibrations throughout the timeframe. Results are shown in Figure 4.14.

As anticipated, the highest values for both MAE and MSE are observed with a single calibration, reflecting the expected trend. However, a notable reduction in error, approximately 9.8 mg/dL for the LMU model and 8.31 mg/dL for the TCN model, is discerned with just three calibrations—equivalent

to one calibration every five days. Interestingly, the error diminishes further to a minimum at five calibrations, corresponding to one calibration every three days. At this point, the models achieve errors of 14.42 mg/dL and 15.18 mg/dL, respectively.

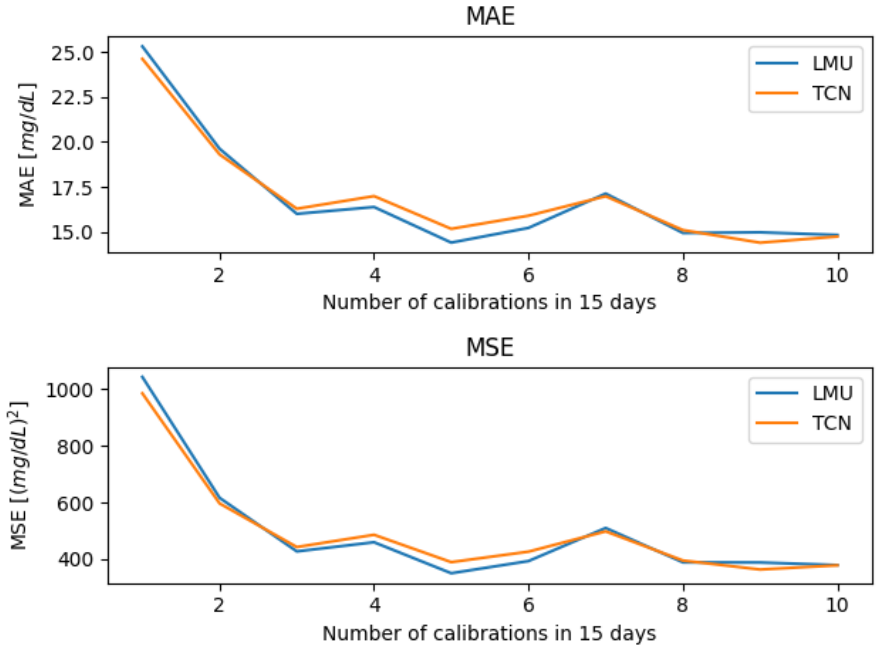


Figure 4.14. Error trends of LMU and TCN models with varying calibration frequency. Plots depict Mean Absolute Error (MAE) and Mean Squared Error (MSE) for both models over 15 days. Error metrics evaluate glucose concentration prediction accuracy, with MAE and MSE.

Assessment of Model Portability for Sensor Deployment

To assess deploying the optimal model on sensors, we used the STM32Cube.AI Developer Cloud tool [289]. The model was compiled using RISC-GCC with size optimisation flags (-Os) and IEEE floating-point format. We implemented Model 11 using ST AI Unified Core Technology in the Intelligent Sensor Processing Unit (ISPU), deploying it on a NUCLEO-F401RE board via SWD interface at 4000 KHz.

The layer-wise performance analysis showed that the third dense layer consumed the majority of inference time (51.5%), followed by the second dense layer (38.5%), while the non-linear activation and first dense layer required only 4.1% and 3.5% respectively. The model achieved a 6.9% reduction in weight memory compared to the original floating-point implementation. Table 4.11 summarizes the key performance metrics.

Table 4.11. Specifications for Model 11: MACCs, flash memory usage, RAM space, and inference time.

MACCs	**Flash size**	**RAM**	**Inference time**
1673	6,47 KiB	320 B	6 ms

Conclusion

In this study, we presented an analysis of the performance of different neural network architectures in predicting the error between a patient's blood glucose level and the sensor reading. The database used to train the model was generated by simulating the responses of 500 different sensors to 15-day interval blood glucose trends of 10 adult patients using the UVA/-Padova simulator.

The 42 models employed four types of architectures: GRU, LSTM, LU, and TCN, with various numbers and types of layers. They were tasked with predicting the time error in the blood glucose level after a single calibration. Two models were selected for their performance: model 11, based on a TCN architecture, achieved the lowest value of MAE (24.22 mg/dL), and model 1, based on LMU, achieved a slightly higher value of MAE (25.34 mg/dL), with a lower number of trainable parameters (only 296). This result is very promising, especially when compared to the error indicated in the datasheets of sensors on the market, which can reach 80 mg/dL depending on sensor usage and environmental conditions.

The performance of these models was then evaluated with different cal-

ibration frequencies ranging from 1 to 10 times over the 15-day period of usage. The results showed a substantial decrease in error when the sensor was calibrated once every 5 days, and a minimum error with the frequency of once every 3 days for both models.

These findings suggest promising applications for improving the accuracy of glucose monitoring systems, which are crucial for effective diabetes management. Future research could explore further optimisation of the selected models and investigate their integration into real-world glucose monitoring devices. Additionally, expanding the dataset to include a more diverse range of patient profiles and sensor types could enhance the generalizability of the models and provide deeper insights into their performance across different contexts.

4.4 Breath Analysis

4.4.1 Background and Fundamentals

Exhaled breath volatile organic compounds (VOCs) consist of thousands of low-weight molecules, some of which originate from cellular metabolism and provide valuable insights into an individual's health status [298]. These compounds reflect biochemical processes occurring within the body and can serve as potential biomarkers for various physiological and pathological conditions. The ability to perform frequent and standardised collection of exhaled breath could enable regular and ubiquitous disease monitoring, significantly enhancing prevention effectiveness [299]. This is particularly relevant in healthcare scenarios where continuous monitoring is crucial but traditional approaches may be overly invasive or burdensome for patients, such as in chronic respiratory conditions, metabolic disorders, and during cancer treatment.

The breath fingerprint, obtained by analysing these VOCs through gas sensor arrays (often called electronic noses), has demonstrated remarkable capabilities, particularly in oncology. Studies have shown that these systems can effectively distinguish lung cancer patients from high-risk current and former smokers with promising sensitivity and specificity [300, 301]. Moreover, recent evidence suggests their potential in monitoring the effectiveness of both surgical and non-surgical therapies in lung cancer treatment, offering a non-invasive approach to therapeutic response assessment [302, 303].

However, the practical implementation of breath analysis technologies faces several challenges that need to be systematically addressed. A critical requirement is the development of user-friendly, standardised, and affordable methods for breath collection that can be deployed in various clinical and home settings. The effectiveness of breath analysis heavily depends on the quality and reproducibility of sample collection, making the sampling methodology a crucial determinant of the technology's success. Addi-

tionally, standardization of analytical procedures and interpretation frameworks is necessary to ensure reliable and comparable results across different research and clinical contexts.

4.4.2 Device set up and pilot test in the longitudinal study of lung cancer

Collection Devices and Methods

Traditional approaches to breath collection often involve complex procedures that can be challenging for both healthcare providers and patients. These methods frequently require specialised equipment and controlled environments, limiting their applicability in routine clinical settings. To address these limitations, recent technological advances have focused on developing more practical and user-friendly collection devices.

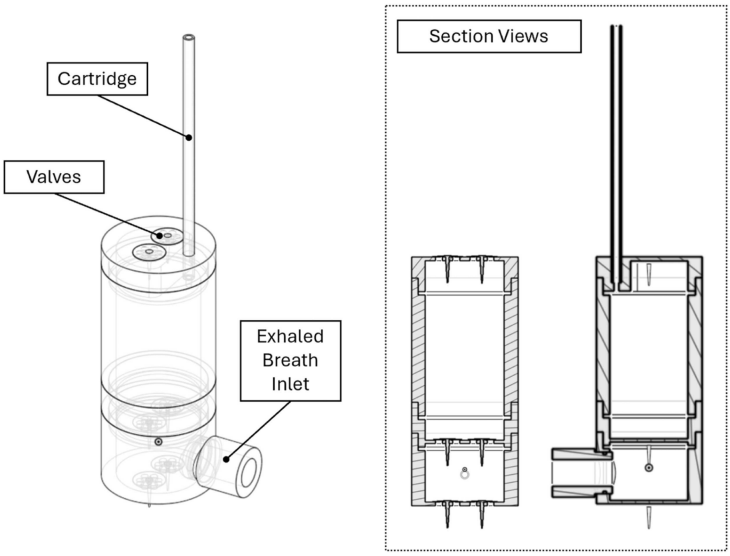


Figure 4.15. Axonometric and longitudinal section view of the Pneumopipe II sampler. The device's double-chamber structure enables continuous sampling of exhaled air, even during inhalation. Key improvements over the previous version include simplified valve configuration and enhanced material compatibility with sterilisation processes.

A significant advancement in this field is the Pneumopipe system (European patent EP2641537 2013), which has evolved into its second iteration, Pneumopipe II. This device represents a crucial step forward in breath collection technology, offering several key improvements over its predecessor:

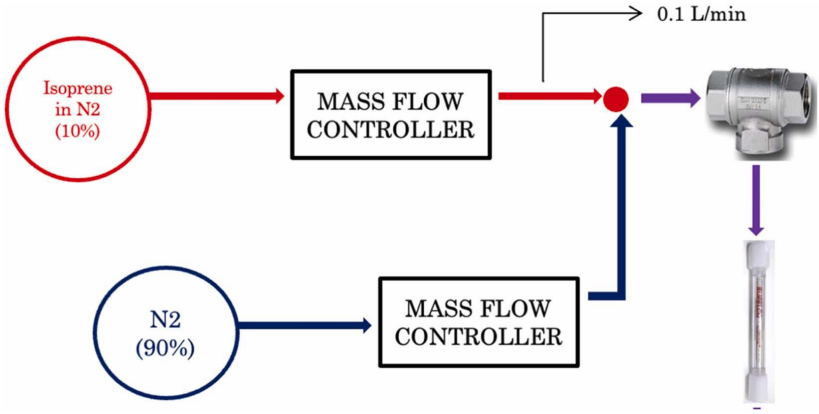
- Simplified design with reduced component count
- Introduction of reusable silicone one-way valves
- Enhanced material compatibility with hydrogen peroxide sterilisation
- Significant cost reduction in both production and operation

The effectiveness of the Pneumopipe II was validated through laboratory testing using isoprene as a reference compound. Isoprene represents an ideal test candidate as it is one of the most abundant endogenous VOCs in human breath and can serve as a biomarker in various conditions. Decreased isoprene levels have been observed in lung cancer patients [304], while increased levels can result from physical exercise [305], making it a valuable compound for system validation.

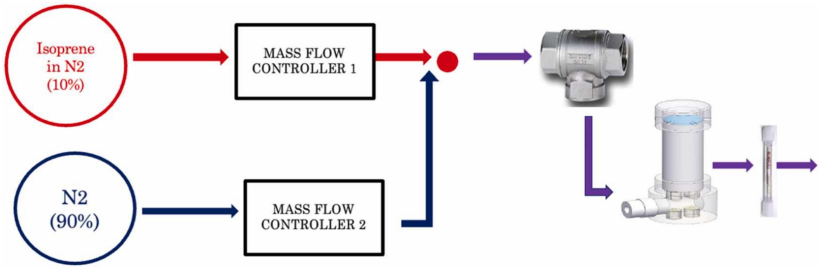
Laboratory validation employed a sophisticated test setup incorporating mass flow controllers for both nitrogen and isoprene channels. The system generated a controlled gas mixture comprising 90% nitrogen and 10% isoprene at a concentration of 100 ppm, resulting in a final isoprene concentration of 10 ppm. Tests were conducted both with and without the Pneumopipe II in the measurement chain to assess its impact on collection efficiency.

Table 4.12. Peak Heights and Areas of GC-FID identification of isoprene in comparative testing

Test Setup	**Peak Area**	**Peak Height**
Setup without Pneumopipe	377,700,029	8,067,803
Setup with Pneumopipe	73,130,822	3,226,88



(a) Direct sampling configuration



(b) Configuration with Pneumopipe II integration

Figure 4.16. Experimental setups for isoprene testing. Both configurations utilize mass flow controllers for precise gas mixture generation. Setup (a) provides baseline measurements while setup (b) validates Pneumopipe II collection efficiency.

Sensing Technologies

The analysis of collected breath samples employs the BIONOTE-V system, where 'V' designates its specialisation in volatiles detection. This advanced sensing platform incorporates an array of eight quartz crystal microbalance (QMB) sensors, each operating at a fundamental frequency of 20 MHz. The sensors feature gold electrodes functionalised with different anthocyanins, enabling selective detection of various volatile compounds.

The sensing mechanism relies on the interaction between VOCs and the functionalised sensor surfaces, which induces measurable shifts in the crys-

tals' fundamental oscillation frequency. This frequency shift is not solely correlated with individual compound concentrations but rather reflects the complex combination of organic compounds present in the breath sample. The system performs thermal desorption analysis at four distinct temperature steps (50°C, 100°C, 150°C, 200°C), generating a comprehensive 32-point response pattern—the breath fingerprint.

Clinical Applications

The integration of the Pneumopipe II collection system with BIONOTE-V analysis has demonstrated particular promise in oncology, specifically in monitoring lung cancer recurrence post-surgery. A pilot study involving 35 patients who underwent surgical treatment for non-small cell lung cancer (NSCLC) provided compelling evidence of the system's clinical utility.

The study protocol involved collecting breath samples at multiple time-points:

- Pre-surgery: Three collections from thirty days to immediately before surgery
- Post-surgery: Three collections within one month after lobectomy

Table 4.13. Study Population Characteristics

Characteristic	Value
Age (years, mean ± SD)	69 ± 8
Male sex	21 (60%)
Current smokers	7 (20%)
Former smokers	21 (60%)
Adenocarcinoma	29 (83%)
Squamous cell carcinoma	3 (9%)
Other histology	2 (8%)
Recurrence during follow-up	8 (23%)

Clinical Validation and Results

The study population included 35 participants with a mean age of 69 years, predominantly male (60%) and with a history of smoking (80%). Adenocarcinoma was the most common NSCLC type (83%). During the one-year follow-up period, eight patients experienced cancer recurrence, with a median time to recurrence of 11 months (range: 3.6–13.8 months).

Analysis of post-surgical breath fingerprints demonstrated remarkable diagnostic capabilities. The system achieved:

Table 4.14. Detection Performance for Cancer Recurrence Based on Post-Surgery Breath Analysis

Performance Metric	**Value**
Accuracy	91%
Sensitivity	75%
Specificity	96%
Positive Predictive Value	86%
Negative Predictive Value	93%

Of particular interest was the analysis of longitudinal breath fingerprint evolution. In a subset of fifteen patients who underwent multiple breath collections both before and after surgery, Principal Component Analysis (PCA) revealed distinct clustering patterns. The first two principal components accounted for more than 75% of breath fingerprint variability, demonstrating clear discrimination between pre- and post-operative states in most cases.

A detailed PCA analysis was performed on fifteen patients who had multiple measurements both before and after surgery. The analysis revealed clear clustering patterns in most cases, with distinct separation between pre- and post-operative measurements. However, in three cases involving recurrence, the post-operative measurements did not form a distinct cluster from the pre-operative ones, suggesting a potential return to cancer-associated breath patterns. This observation further supports the system's

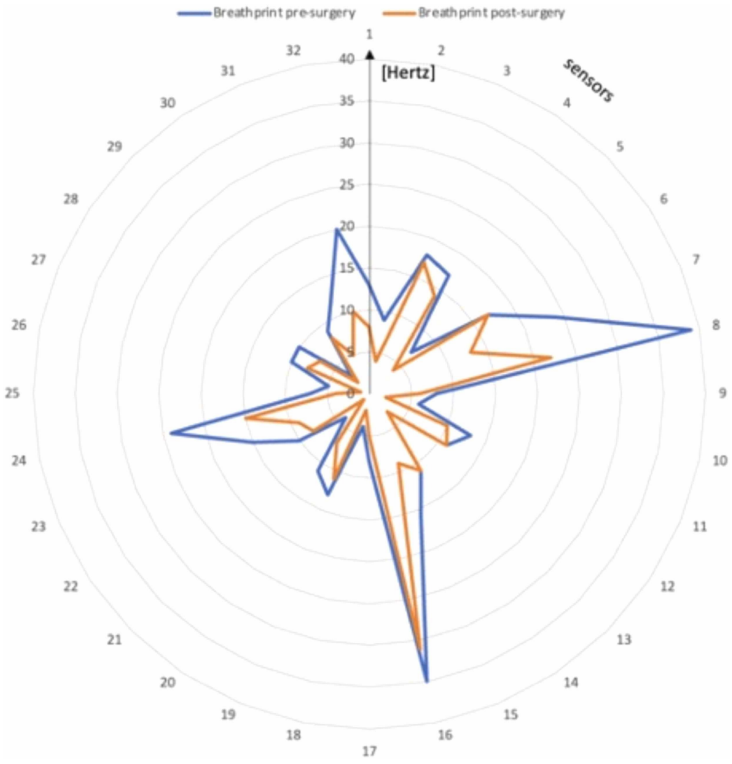


Figure 4.17. Representative radar plot of breath fingerprint patterns before (blue) and after (orange) surgery, demonstrating distinct changes in VOC profiles.

capability to detect disease recurrence through breath pattern analysis.

Future Directions

The integration of Pneumopipe II collection system with BIONOTE-V analysis presents promising opportunities for reshaping post-surgical monitoring protocols in lung cancer patients. The system's ability to detect recurrence with high accuracy, combined with its non-invasive nature and relatively low cost (approximately 10 euros per measurement), makes it an attractive complement to current follow-up methods.

Several areas warrant further investigation:

- Validation in larger patient populations to confirm preliminary findings
- Investigation of breath pattern variations across different cancer histotypes
- Development of standardised libraries for breath pattern comparison
- Integration with existing follow-up protocols to optimise early detection of recurrence

The technology's potential extends beyond lung cancer surveillance. Similar approaches could be valuable in monitoring response to non-surgical treatments such as chemotherapy and radiotherapy. Furthermore, the temperature-dependent VOC extraction process might enable identification of specific sentinel molecules associated with disease recurrence.

While these results are promising, certain limitations must be acknowledged. The current pilot study's sample size, while sufficient for preliminary validation, necessitates larger confirmatory studies. Additionally, findings are specific to the BIONOTE-V system, and generalisability to other electronic nose technologies requires further investigation.

4.5 Urine Culture Spectrophotometry

4.5.1 Background

Bladder cancer poses a substantial global health challenge [306, 307], with significant impact on patients' quality of life [308]. While various screening methods exist for early detection of high-risk patients, there is currently no standardised routine screening test [309]. The conventional diagnostic approaches have significant limitations - cystoscopy is invasive, while urinary cytology is characterised by a high percentage of false negative results [310]. The need for non-invasive diagnostic approaches is particularly relevant given the global burden of bladder cancer, with over 573,278 new cases diagnosed worldwide in 2020 [307]. This number is expected to double by 2040 based on World Health organisation predictions [311]. Early detection is crucial as it signifies better prognosis, highlighting the importance of developing minimally invasive diagnostic options to improve patient outcomes. Recent advances in optical sensing technologies and our understanding of tissue-light interactions have enabled the development of novel diagnostic approaches [310]. Spectrophotometric methods analyse the optical properties of urine samples using wavelengths between 340-850 nm, allowing for non-invasive detection of bladder cancer markers. The underlying principle relies on the distinct spectral characteristics that arise from alterations in cellular and molecular components associated with malignant transformation. The development of compact spectrometers represents a promising direction in diagnostic technology, offering potential advantages in terms of speed, cost-effectiveness and accessibility compared to conventional methods [312]. However, optimal implementation requires careful consideration of both technical performance and clinical utility, particularly regarding sensitivity and specificity for early-stage disease detection.

4.5.2 Study on predicting urine culture outcome via spectrophotometry

Methods and Technical Principles

The development of this novel compact spectrometer for bladder cancer detection was based on a systematic approach to non-invasive spectrophotometric diagnostics. Between January 2022 and July 2023, 300 patients who underwent transurethral resection of the bladder (TURB) were enrolled in the study. A urine sample was collected from each patient before surgery for spectrophotometric analysis. The measurement system comprises a compact spectrometer with an operating range between 340 and 850 nm. The device's optical configuration includes:

- A miniaturised spectrometer unit
- An LED light source
- A standardised urine sample holder

From the original 288 features within the spectrogram data, the initial 256 features were selected, enabling the data to be processed as a 16×16 pixel image. This feature selection process was designed to ensure that discarded features contained negligible pertinent information while maintaining the diagnostic value of the spectral data. The measurement protocol was standardised to ensure reproducible results:

1. Collection of urine samples in standardised vacutainers
2. Sample analysis using the compact spectrometer under controlled conditions
3. Data acquisition and processing using standardised parameters

Data Processing and Analysis

The data analysis methodology incorporated multiple stages to ensure robust classification of the spectrophotometric data. Key processing steps

included feature extraction and machine learning classification. From the urine spectrogram data, several analytical approaches were implemented:

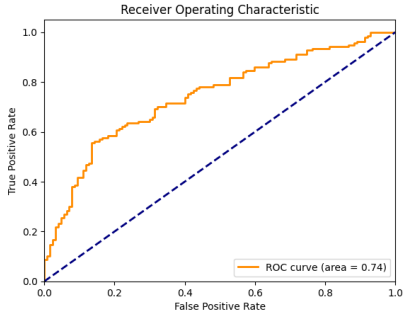
- Extraction of histograms
- Computation of entropy and energy parameters
- Haralick feature extraction from spectral images

The dataset was divided into training and testing sets using Stratified K-Fold cross-validation. For certain cases where class imbalance was present, undersampling was applied. Sequential Forward Selection was employed for feature selection, and hyperparameter optimisation was conducted via Grid Search CV. Multiple machine learning algorithms were evaluated for classification performance, specifically kNN, Random Forest, SVM, and Gradient Boosting Classifier. The evaluation metrics used to assess model performance included Accuracy, Precision, Recall and F1-score.

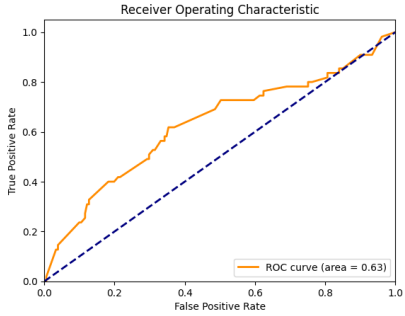
Results

The spectrophotometric analysis demonstrated varying levels of diagnostic accuracy across different classification tasks. Performance evaluation included receiver operating characteristic (ROC) curves analysis, accuracy metrics, and precision-recall evaluations for each classification approach.

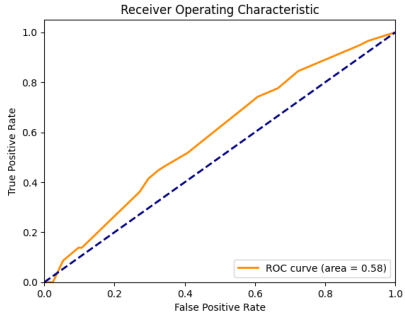
The best performance was achieved in histological grading, where the Gradient Boosting Classifier demonstrated excellent discrimination (AUC = 0.80, accuracy = 80%, precision = 0.81). This classifier proved particularly effective in distinguishing between high and low-grade samples, suggesting robust capability for tumour grading. Cytological assessment using the Random Forest model showed good discriminative ability (AUC = 0.74) with moderate accuracy (67%) and precision (0.38). While less accurate than histological grading, this performance level suggests potential utility as a screening tool. The kNN model for urine culture classification achieved modest results (AUC = 0.58, accuracy = 65%, precision = 0.27), indicating limited discriminative power for this specific application. Basic



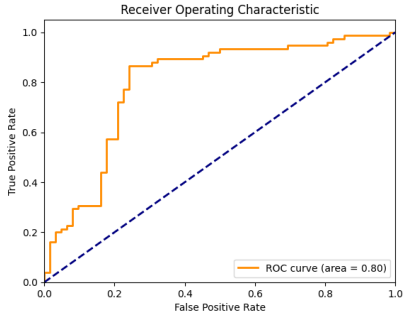
(a) ROC curve for histological grading (high/low), AUC = 0.80



(b) ROC curve for cytological assessment, AUC = 0.74



(c) ROC curve for urine culture classification, AUC = 0.58



(d) ROC curve for basic histological assessment, AUC = 0.63

Figure 4.18. ROC curves for different classification tasks using spectrophotometric analysis.

histological assessment using the Gradient Boosting Classifier showed intermediate performance (AUC = 0.63), though with notably high precision (0.79) for positive cases. Table 4.15 summarises the key performance metrics across all classification tasks:

These results demonstrate that spectrophotometric analysis shows particular promise for histological grading, while its utility for urine culture classification may require further optimisation or complementary diagnostic approaches.

Table 4.15. Performance metrics for different classification tasks

Classification Task	Accuracy (%)	Precision	F1-score	AUC
Histological Grading	80	0.81	82	0.80
Cytological Assessment	67	0.38	47	0.74
Urine Culture	65	0.27	30	0.58
Basic Histological	70	0.79	67	0.63

Discussion

The results of this study demonstrate that spectrophotometric analysis using a compact spectrometer can effectively detect bladder cancer in human urine samples, particularly for histological grading applications. The varying performance across different classification tasks provides important insights into both the capabilities and limitations of this technology. The superior performance in histological grading (AUC = 0.80) suggests that the spectral signatures captured by the system are particularly sensitive to the morphological and biochemical changes associated with tumour grade differentiation. This finding aligns with previous studies showing that optical properties of tissue correlate strongly with histopathological features [309]. While the cytological assessment showed moderate performance (AUC = 0.74), it represents a significant advancement in automated urine sample analysis. The lower precision (0.38) compared to conventional cytology may be offset by the advantages of rapid, automated processing and the elimination of inter-observer variability. However, this suggests that the technology might be better positioned as a screening tool rather than a definitive diagnostic test. The relatively modest performance in urine culture classification (AUC = 0.58) highlights current limitations in detecting bacterial presence through spectrophotometric means alone. This may be due to:

- The complexity of urine composition and its effects on spectral measurements

- Possible interference from non-bacterial elements in the samples
- The challenge of detecting low bacterial concentrations using current spectral resolution

A notable finding is the high precision (0.79) achieved in basic histological assessment for positive cases, despite a moderate overall AUC of 0.63. This suggests that while the system may have a higher false-negative rate, it demonstrates good reliability in positive case identification. The compact nature of the spectrometer and its ability to analyse samples without complex preparation represents a significant advancement toward more accessible diagnostic tools. However, several technical and clinical considerations must be addressed:

1. Optimisation of the wavelength range for specific diagnostic tasks
2. Development of more robust classification algorithms
3. Validation in larger, multi-centre clinical trials
4. Integration into existing clinical workflows

Discussion and Conclusions

The results presented in this study demonstrate the feasibility and efficacy of using a compact spectrophotometer for urine analysis to enable early detection of bladder cancer. This novel non-invasive diagnostic tool has shown particular promise in discriminating high-grade bladder tumours, with performance metrics suggesting strong potential for clinical utility. The spectrophotometric approach was able to achieve 80% accuracy in histological grading of tumours, with especially high precision of 0.81 for detecting high-grade cancers. This indicates that the technology could offer an effective way to rapidly and non-invasively screen for clinically significant bladder tumours while maintaining good overall diagnostic accuracy. Notably, in some cases the spectrophotometric system even outperformed standard histopathological evaluation by human pathologists in predicting

high-grade cancers. While these initial results are promising, further work is needed to optimise and validate the technology before clinical implementation. From a technical perspective, next steps could include fine-tuning the analysed wavelength range for specific diagnostic applications, enhancing the signal processing algorithms to improve sensitivity and specificity, and developing more robust feature extraction methods. Clinically, larger multicentre trials and longitudinal studies are necessary to rigorously assess the long-term reliability and reproducibility of this diagnostic approach in diverse patient populations. The spectrophotometric method should also be benchmarked against emerging molecular biomarker-based tests for bladder cancer to compare their relative performance, cost-effectiveness, and clinical utility. Successful translation of this technology into clinical practice will require careful consideration of how to optimally integrate it into existing diagnostic pathways for bladder cancer. This includes developing standardized protocols for urine sample collection and spectrophotometric analysis, as well as performing health economic evaluations to establish the cost-effectiveness of this approach in various healthcare settings and populations. By reducing the need for invasive cystoscopy procedures, this new technology could potentially decrease both patient burden and healthcare costs while maintaining high diagnostic standards if implemented properly. In conclusion, the compact spectrophotometer platform described here represents a promising step towards improving the accessibility, efficiency and non-invasiveness of bladder cancer diagnostics. With further refinement and validation, this novel technology could significantly enhance capabilities for early detection and monitoring of bladder tumours, thereby facilitating more timely intervention and ultimately improving patient outcomes. While technical and translational challenges remain, the successful development and implementation of this spectrophotometric approach could revolutionize the clinical pathway for bladder cancer diagnostics.

4.6 Integration with Human-centred Design Principles

The sensing technologies and calibration approaches presented in this chapter fundamentally align with the human-centred design principles outlined in Chapter 1, where we emphasized the development of unobtrusive technological solutions that enhance quality of life while maintaining natural human interaction patterns. This section explicitly examines how each research area contributes to these principles and embodies the core philosophy of user-centred healthcare monitoring.

4.6.1 Human-centred Aspects of Glucose Sensor Calibration

The innovations in glucose sensor calibration directly address several key human-centred design challenges identified in the introduction. By developing novel calibration strategies that reduce user intervention requirements, this research specifically tackles what Nthubu et al. [243] identified as critical factors for technology adoption: minimal disruption to normal activities and reduced cognitive load.

The optimisation of neural network architectures for glucose sensing represents a concrete implementation of the ergonomic design principles discussed by Pickham et al. [244]. By creating lightweight models capable of running on resource-constrained devices, this research enables:

- **Reduced Calibration Frequency:** As demonstrated in our findings, the optimised models can maintain acceptable accuracy with only three calibrations over 15 days—equivalent to one calibration every five days. This directly reduces the user burden compared to conventional systems requiring daily calibration.
- **Enhanced User Independence:** The ability to implement these algorithms directly on glucose sensors creates the potential for more

autonomous systems that require less conscious attention from users, addressing the core principle of minimizing cognitive load.

- **Contextual Adaptability:** By modelling various interfering factors such as temperature, pH, and adhesion, our approach enables sensors to automatically compensate for the diverse real-world conditions users experience, rather than requiring users to adapt to the technology's limitations.

The shift from universally applied calibration schedules to personalized, adaptive approaches represents a fundamental reorientation toward user-centred design. Rather than imposing a technological routine on users, these systems adapt to individual physiological characteristics and lifestyle patterns—embodying what Ming et al. [240] described as the evolution from technology-centred to human-centred healthcare monitoring.

4.6.2 User-centred Principles in Breath Analysis

The breath analysis technology presented in Section 4.4 exemplifies human-centred design through its focus on creating accessible, non-invasive monitoring solutions. The Pneumopipe II collection system specifically addresses several human factors identified as critical for successful implementation:

- **Simplified User Interaction:** The redesigned device reduces component count and complexity, creating a more intuitive user experience that requires minimal training—directly addressing the issue of cognitive load highlighted by Nthubu et al. [243].

- **Enhanced Affordability:** By reducing production costs to approximately 10 euros per measurement, the system increases accessibility across diverse socioeconomic contexts, addressing the socioeconomic barriers to healthcare monitoring technologies identified in our introduction.

- **Integration with Existing Clinical Workflows:** The system's design enables seamless integration into established clinical protocols, minimizing disruption to healthcare providers' routines while maximizing potential adoption.

The development of standardized collection protocols directly addresses what Wang et al. [246] identified as a critical need for reproducible methodologies in healthcare sensing. By ensuring consistent sampling quality across users and settings, the Pneumopipe II enhances result reliability without increasing user burden.

The promising results in cancer recurrence detection (91% accuracy) demonstrate the potential of this human-centred approach to positively impact clinical outcomes. This aligns with our fundamental thesis objective of developing technologies that enhance quality of life while maintaining natural human interaction patterns.

4.6.3 Human-centred Aspects of Spectrophotometric Analysis

The spectrophotometric approach to urine analysis exemplifies several core principles of human-centred sensing:

- **Minimal User Intervention:** By using standard urine samples without complex preparation requirements, the system minimizes the procedural burden on both patients and healthcare providers—addressing a key barrier to adoption identified by Ates et al. [245].

- **Rapid Analysis:** The system's ability to provide immediate results contrasts sharply with conventional pathology processing times, potentially reducing patient anxiety and enabling more timely clinical decision-making.

- **Non-Invasive Monitoring:** By potentially reducing the need for invasive cystoscopy procedures, this approach directly addresses the prin-

ciple of minimizing physical discomfort while maintaining diagnostic effectiveness.

The spectrophotometric system's compact design and operational simplicity align with what Wasilewski et al. [248] identified as critical for successful integration into clinical practice. By requiring minimal technical expertise to operate, the system reduces implementation barriers and enhances potential adoption across diverse healthcare settings.

4.6.4 Cross-Cutting Human-centred Design Elements

Several human-centred design elements transcend individual applications and represent common threads throughout this chapter:

- **Privacy-Preserving Processing:** By implementing localized processing (as in the glucose sensing application) or aggregated feature extraction (in the spectrophotometric analysis), these approaches limit the transmission of raw health data, addressing the privacy concerns identified by Karalis et al. [251].

- **User Autonomy Enhancement:** Each technology reduces dependency on specialized healthcare facilities for routine monitoring, potentially enabling greater patient autonomy and self-management—a core principle identified by Iqbal et al. [252].

- **Reduced Healthcare Burdens:** By potentially decreasing the frequency of clinical visits required for monitoring (glucose sensing), providing early warning of recurrence (breath analysis), or eliminating the need for invasive procedures (spectrophotometry), these technologies directly address healthcare system burdens while improving patient experience.

4.6.5 Limitations and Future Human-centred Design Opportunities

While the technologies presented demonstrate significant progress toward human-centred sensing, several limitations and future opportunities merit consideration:

- **Cultural and Contextual Adaptation:** As noted by Bhaltadak et al. [241], successful implementation of health technologies requires adaptation to diverse cultural contexts and user expectations. Future work should examine how these technologies perform across more diverse populations and contexts.
- **Participatory Design Enhancement:** While these technologies address identified user needs, future iterations would benefit from more systematic inclusion of end-users in the design process, as recommended by Nthubu et al. [243].
- **Long-Term Usage Considerations:** As Spring et al. [249] note, healthcare sensing must consider not just initial adoption but sustainable long-term usage patterns. Future research should investigate how user experience evolves over extended periods of use for these technologies.

These considerations highlight the ongoing nature of human-centred design as an iterative process rather than a fixed achievement. The technologies presented represent significant steps toward more human-centred healthcare sensing, while acknowledging the need for continued refinement based on real-world implementation experiences.

In summary, the sensing technologies and calibration approaches presented in this chapter demonstrate concrete implementations of the human-centred design principles outlined in our introduction. By prioritizing user needs, minimizing intervention requirements, and enhancing accessibility, these approaches advance our fundamental research objective: developing

technologies that enhance human capabilities while minimizing their impact on natural behaviour and experience.

Chapter 5

Conclusions

Quayola, "Storms #3" (2020), digital print from algorithmic simulation, 160 × 90 cm.

5.1 Key Contributions

This thesis presents a series of interconnected contributions spanning the domains of human-technology interaction, affective computing, and healthcare sensing technologies. The work is unified by a fundamental focus on enhancing the integration of emotional awareness and human factors in healthcare technology development. This approach recognises that successful technological innovation in healthcare requires not only technical excellence but also careful consideration of the human experience, including emotional states, cognitive processes, and behavioural patterns that influence technology adoption and efficacy.

Throughout the research presented, several unifying themes emerge. First, the integration of emotional awareness into technological solutions has proven crucial for improving both user acceptance and clinical outcomes. This is evident across diverse applications, from sensory substitution devices to music therapy systems, where emotional congruence significantly enhances user engagement and therapeutic efficacy. Second, the importance of human-centred design principles has been demonstrated consistently, particularly in the development of non-invasive monitoring solutions that prioritise user comfort and natural interaction patterns, thereby reducing the burden of healthcare monitoring while maintaining clinical utility. Third, the value of quantitative validation approaches in emotionally-aware technologies has been established, providing robust frameworks for assessing both technical performance and emotional congruence, enabling evidence-based refinement of these systems.

The research adopts a multidisciplinary approach, combining insights from psychology, neuroscience, and engineering to address complex healthcare challenges. This integration of perspectives has enabled the development of more comprehensive and effective solutions, as demonstrated in the various studies presented. The collaborative methodology employed throughout this work exemplifies how cross-disciplinary research can overcome traditional barriers between technical and human-centred approaches.

The following sections detail the specific contributions in each research area, highlighting their immediate impact and broader implications for healthcare technology development.

5.1.1 Understanding Emotion in Human-Technology Interaction

In the domain of emotion-technology interaction, this thesis presents several significant advances in understanding how emotional responses mediate human perception and interaction with technological systems. The investigation into crossmodal perception, particularly focusing on audiovisual associations, has yielded important insights for the development of more intuitive and emotionally congruent interfaces that align with users' natural sensory expectations.

The systematic study of colour-sound mapping in children and adults (Chapter 2.4) revealed distinct patterns in how different age groups process and associate sensory information across modalities. The research demonstrated consistent mapping preferences between musical parameters and colour properties, while also highlighting important developmental differences in these associations that reflect cognitive maturation processes. The methodology developed for this study provides a robust framework for quantifying emotional responses in audiovisual interactions, with broader applications in interface design and assistive technology development, particularly for age-appropriate technological interventions.

Building on these findings, the experimental validation of emotion-mediated audiovisual associations (Chapter 2.5) provided quantitative evidence for the role of emotions in crossmodal perception and multimodal integration. This work established a comprehensive protocol for measuring emotional mediation in sensory associations, supported by statistical validation of the relationship between emotional responses and sensory parameter selection across diverse stimuli. The resulting framework enables prediction of appropriate sensory mappings based on emotional content, advancing our un-

derstanding of how emotions influence human-technology interaction and providing practical guidelines for designing more emotionally congruent interfaces that resonate with users' perceptual expectations.

5.1.2 Advances in Affective Computing

The research in affective computing has produced three major contributions, each addressing critical challenges in the integration of emotional awareness into healthcare technologies. The development of the Universal Validation Protocol for Facial Expression Recognition (FER) algorithms and the creation of the FeelPix database (Chapter 3.5) represents a significant methodological advance in the field. This work established standardised evaluation procedures for FER algorithms while creating a comprehensive database incorporating self-reported emotional states alongside facial expression data. The emphasis on lightweight implementation strategies ensures applicability across a range of devices, including resource-constrained systems, thereby expanding the potential deployment scenarios for emotion-aware healthcare applications.

The AFFECT-SENSE system (Chapter 3.6) explores the theoretical and practical foundations for integrating emotional awareness into sensory substitution devices for individuals with sensory impairments. The research investigates how emotional responses influence audiovisual associations through extensive experimental studies and data analysis across different user groups. The development of predictive models for emotional responses in audiovisual mappings, coupled with the implementation of lightweight neural architectures suitable for embedded systems, provides a framework for future emotion-aware sensory substitution devices that enhance user experience. This work advances our understanding of how emotional congruence can be maintained in sensory substitution while establishing practical guidelines for implementing such systems in resource-constrained environments with minimal computational overhead.

Applications in music therapy have provided valuable insights into the

clinical utility of affective computing in therapeutic contexts. The implementation of emotion recognition systems in therapeutic settings, combined with validation of technology-enhanced music therapy interventions, has demonstrated improved patient engagement and therapeutic outcomes across various clinical populations. These findings have important implications for the future development of technology-supported therapeutic interventions, particularly in mental health and rehabilitation settings where emotional awareness is central to treatment efficacy.

5.1.3 Innovations in Healthcare Sensing

The research presents significant advances in healthcare sensing technology, particularly in the areas of glucose monitoring, breath analysis, and spectrophotometric diagnostics. In glucose sensor calibration (Chapter 4.3), the development of a comprehensive model for sensor interference effects, coupled with the creation of a synthetic dataset for algorithm validation, has enabled more effective calibration strategies for continuous monitoring systems. The implementation of efficient neural network architectures for real-time calibration represents a practical solution for improving continuous glucose monitoring accuracy while reducing calibration requirements and patient burden, addressing key limitations in current diabetes management technologies.

The breath analysis research (Chapter 4.4) has advanced non-invasive diagnostic capabilities through validation of the Pneumopipe II collection system and its integration with the BIONOTE-V analysis platform for volatile organic compound detection. Clinical validation studies in lung cancer monitoring have demonstrated the technology's potential for disease surveillance and treatment monitoring, establishing a foundation for broader applications in respiratory diagnostics and personalized medicine approaches. The standardized collection methodology developed addresses critical challenges in sample reproducibility that have limited previous breath analysis technologies.

Advances in spectrophotometric analysis (Chapter 4.5) have resulted in the development of a compact spectrometer for urine analysis, supported by sophisticated machine learning algorithms for diagnostic classification and biomarker identification. Clinical validation in bladder cancer detection has demonstrated the potential for non-invasive diagnostic tools in oncology, with promising implications for both screening and monitoring applications in resource-limited settings. The portable nature of the developed system enables point-of-care diagnostics with minimal infrastructure requirements, potentially expanding access to cancer screening in underserved populations.

5.2 Future Research Directions

The contributions presented in this thesis open several promising avenues for future research. Technical developments should focus on the integration of multiple sensing modalities with affective computing systems, creating multimodal platforms that capture a more comprehensive picture of human physiological and emotional states. Continued enhancement of real-time processing capabilities for emotion-aware systems will be necessary, particularly through edge computing approaches that minimize latency in critical applications. Particular attention should be paid to developing more efficient algorithms suitable for resource-constrained devices, enabling broader deployment of these technologies in clinical settings and home healthcare environments where computational resources may be limited.

Clinical applications represent another crucial area for future research. The expansion of validation studies across different patient populations and healthcare contexts will be essential for establishing the robustness and generalisability of these approaches in real-world settings. Investigation of new therapeutic applications for emotion-aware technologies, particularly in chronic disease management and mental health interventions, coupled with the development of standardised protocols for technology-enhanced interventions, will help realise the full potential of these innovations in healthcare delivery. Longitudinal studies examining the long-term impact of emotionally-aware technologies on patient outcomes and adherence will provide valuable insights for healthcare implementation strategies.

Methodological advances will also be critical for future development. Refinement of validation protocols for emotion-aware technologies, coupled with improved metrics for measuring emotional congruence across diverse user groups, will enable more rigorous evaluation of these systems throughout the development lifecycle. The development of standardised frameworks for evaluating human-technology interaction in healthcare contexts, incorporating both objective performance measures and subjective

user experience, will support more effective translation of research findings into clinical practice. Ethical considerations, particularly regarding privacy, data ownership, and algorithmic bias in emotion recognition systems, will require continued attention as these technologies become more pervasive in healthcare settings.

The convergence of these research directions promises to further advance the integration of emotional awareness in healthcare technology, potentially leading to more effective and user-centred healthcare solutions that address both clinical and human needs. Successful pursuit of these directions will require continued collaboration across disciplines and careful attention to both technical excellence and human factors in healthcare technology development, ultimately creating systems that not only monitor and treat but also empathize and adapt to individual emotional states and preferences.

Appendix A

Appendix A - Interfaces presented in the ASSISI experiment

A.1 Form

In this section the form is reported in Italian as it was presented to the trial subjects.

Informazioni di base

Rispondi a queste domande prima di iniziare il test.

* Indica una domanda obbligatoria

1. Email *

2. Nome e Cognome

3. Data di nascita *

Esempio: 7 gennaio 2019

4. Sesso *

Contrassegna solo un ovale.

- ◯ Uomo
- ◯ Donna
- ◯ Preferisco non rispondere

5. Nazionalità *

6. Con che mano utilizzi il mouse? *

Contrassegna solo un ovale.

- ◯ Destra
- ◯ Sinistra
- ◯ Non uso sempre la stessa mano

Background musicale

7. In media, quanto tempo ascolti brani quotidianamente? *

Contrassegna solo un ovale.

- ◯ Meno di 20 minuti
- ◯ 20 minuti - 1 ora
- ◯ 1-3 ore
- ◯ Più di 3 ore

Indica su una scala da 1 a 10 il tuo gradimento nei confronti dei diversi generi musicali:

8. Classico *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

9. Pop *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

10. Rock *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

11. Jazz e Blues *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

12. Hip hop e rap *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

13. Country-Western *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

14. Elettronica *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

15. Colonne sonore *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

16. Hai mai partecipato a delle lezioni di musica? *

Contrassegna solo un ovale.

◯ Sì *Passa alla domanda 17.*

◯ No *Passa alla domanda 20.*

Background musicale

17. Per quanti anni hai studiato musica? *

Contrassegna solo un ovale.

	0	1	2	3	4	5	6	7	8	9	10	
Men	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Per 10 o più anni

18. Quale/i strumento/i hai studiato? *

_____

19. Quando hai partecipato all'ultima lezione? *

Contrassegna solo un ovale.

- ◯ Meno di 1 anno fa
- ◯ Meno di 2 anni fa
- ◯ Meno di 5 anni fa
- ◯ Meno di 10 anni fa
- ◯ Più di 10 anni fa

Background artistico

In una scala da 1 a 10, indica quanto trovi emozionanti questi colori e forme:

20. Rosso *

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	8	9	10
	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

21. Giallo *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

22. Blu *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

23. Arancione *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

24. Verde *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

25. Viola *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

26. Bianco *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

27. Nero *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

28. Grigio *

Contrassegna solo un ovale.

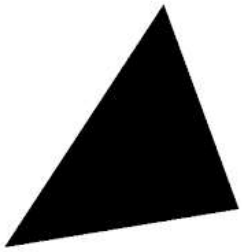
1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

29. Esagono *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

30. Triangolo *



Contrassegna solo un ovale.

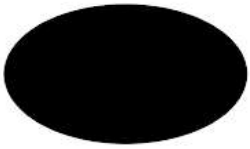
1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

31. Quadrato *

Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

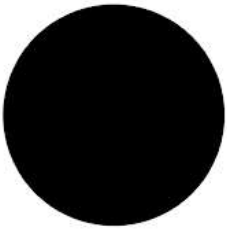
32. Ellisse *



Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

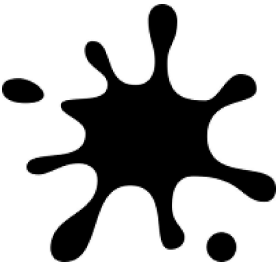
33. Cerchio *



Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

34. Splash *



Contrassegna solo un ovale.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

35. Hai mai partecipato a lezioni di storia dell'arte, pittura, disegno, scultura o simili? *

Contrassegna solo un ovale.

- ◯ Sì *Passa alla domanda 36.*
- ◯ No *Passa alla domanda 39.*

Background artistico

36. A che genere di lezioni hai partecipato? *

_____

37. Per quanti anni? *

Contrassegna solo un ovale.

	0	1	2	3	4	5	6	7	8	9	10	
Men	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Da 10 o più anni

38. Quando hai partecipato all'ultima lezione? *

Contrassegna solo un ovale.

- ◯ Meno di 1 anno fa
- ◯ Meno di 2 anni fa
- ◯ Meno di 5 anni fa
- ◯ Meno di 10 anni fa
- ◯ Più di 10 anni fa

Background medico

39. Ti sono mai stati diagnosticati disturbi visivi? *

Contrassegna solo un ovale.

- ◯ Sì
- ◯ No
- ◯ Preferisco non rispondere

40. Se sì, che tipo di disturbo visivo?

41. Ti sono mai stati diagnosticati disturbi uditivi? *

Contrassegna solo un ovale.

- Sì
- No
- Preferisco non rispondere

42. Se sì, che tipo di disturbo uditivo?

43. Ha mai sperimentato l'esperienza della sinestesia (condizione per la quale quando si subisce uno stimolo, se ne prova un secondo in contemporanea. Ad esempio: si vede un colore quando si sente una musica, si vede una forma se si sente un sapore, …)? *

A.2 Pretest

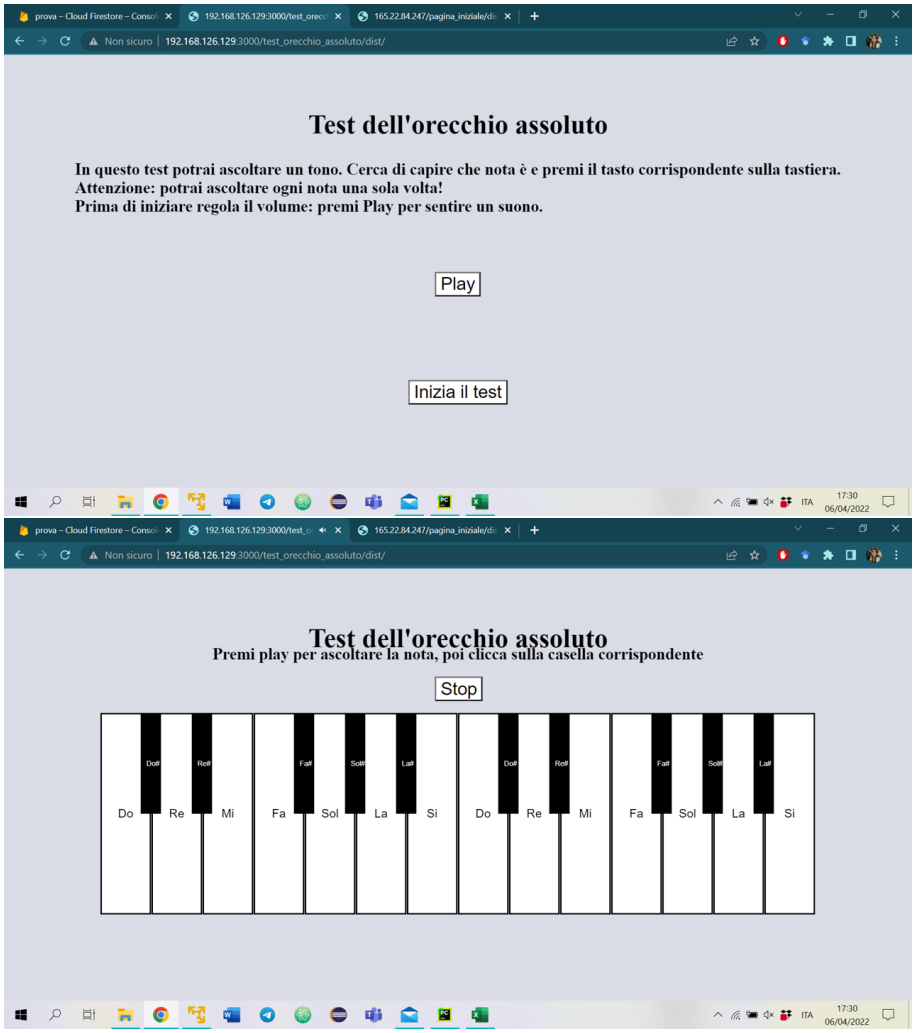


Figure A.1. Perfect Pitch Test: subjects are shown the instructions, asked to play a note and then identify it on the keyboard. They are not allowed to play each note more than once.

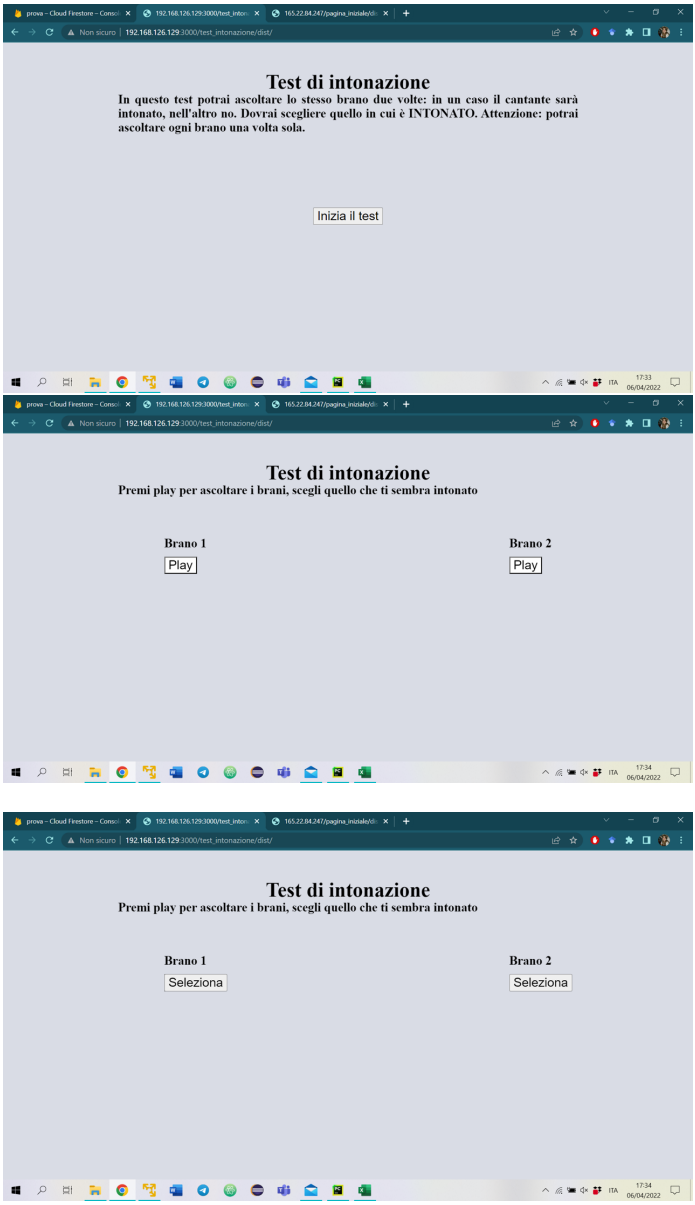


Figure A.2. Mistuning Perception Test: subjects are given instructions to play the two songs in whichever order they prefer and then choose the one where the singer is in tune. They are not allowed to play the songs again.

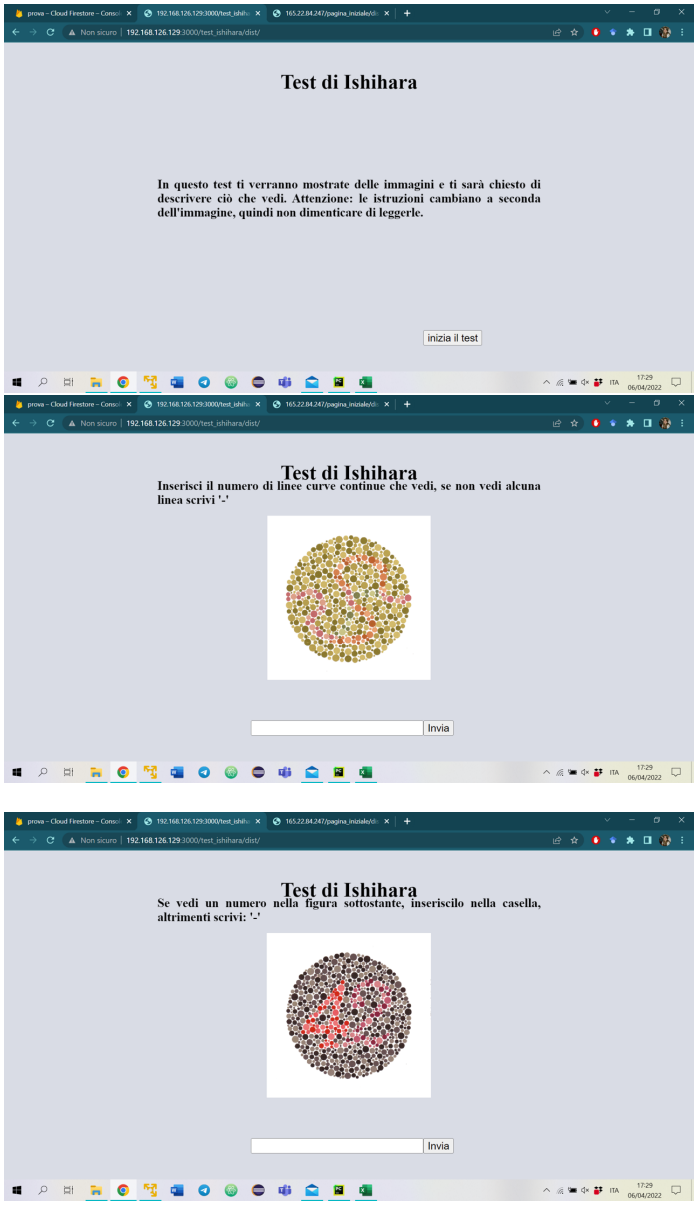


Figure A.3. Ishihara Test: subjects are given instructions and then asked to write the digits or the number of continuous curved line they see in the image.

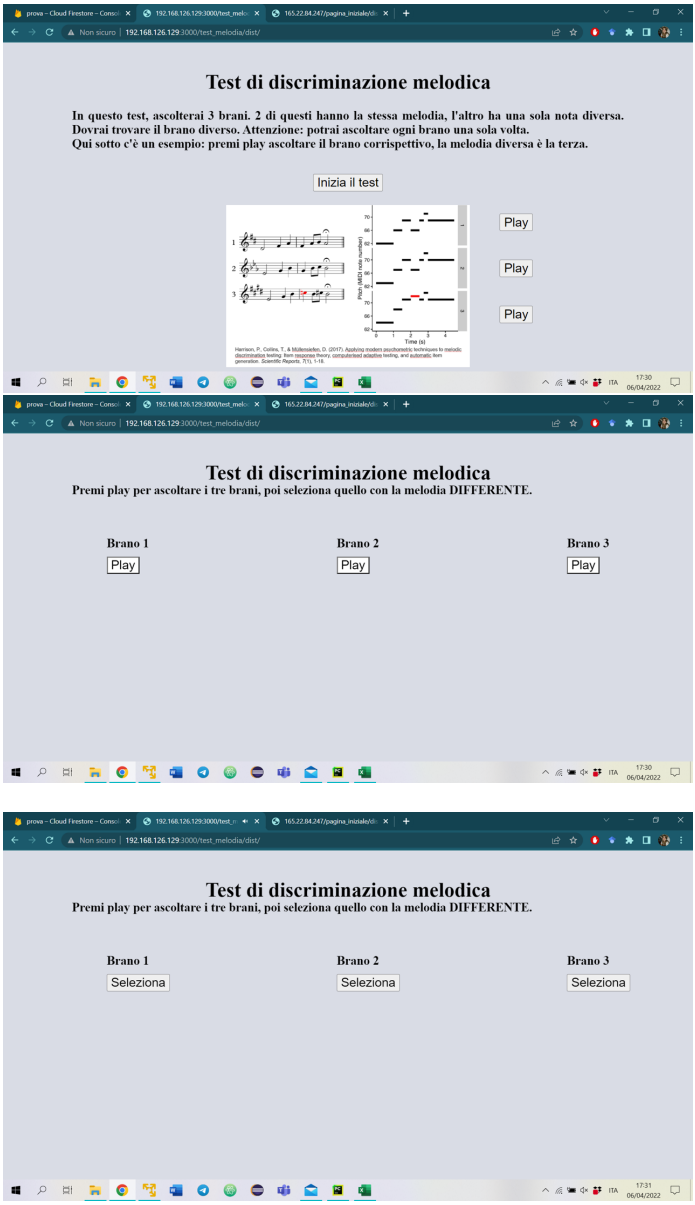


Figure A.4. Melodic Discrimination Test: subjects are given instructions to listen to all three songs once in whichever order they prefer and then select the different melody.

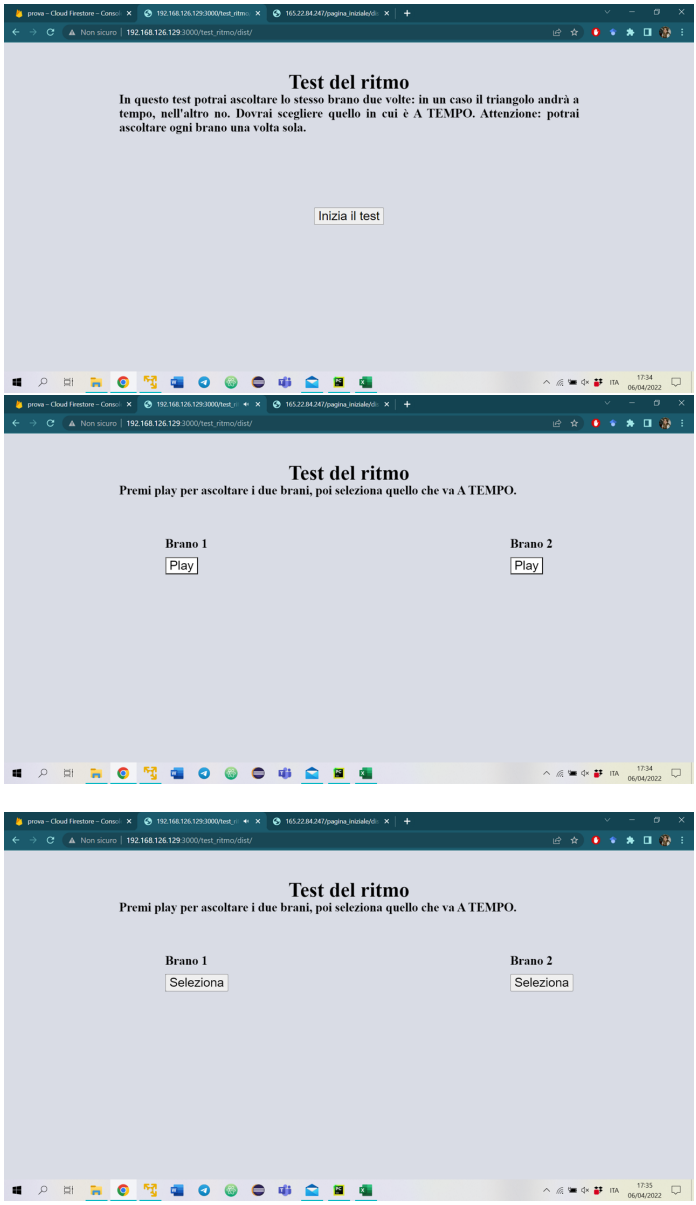


Figure A.5. Beat Alignment Test: subjects are given instructions to listen to the songs once and select the one with the correctly aligned beat.

Appendix B

Appendix B - Explorations in Art and Perception

B.1 Art Turing Test

Introduction

Artificial intelligence (AI) has been increasingly encroaching on creative territories traditionally dominated by humans. Its applications span a multitude of creative fields, and the art world experiences significant transformations due to AI interventions [313, 314]. One remarkable domain where AI has made substantial inroads is generative art: a practice where artists use algorithms and AI to co-create or generate unique works [315, 316].

Generative art has been around for several decades, but the advent of AI has dramatically expanded its possibilities. AI can generate intricate designs, novel patterns, and complex compositions, sometimes achieving levels of creativity that rival human artists [317, 318, 319]. The history of AI art can be traced back to the 1960s, when pioneers such as Harold Cohen and Vera Molnar began exploring the potential of computers for artistic creation [320, 321]. Cohen's AARON and Molnar's Generative Compositions, both algorithm-based art generators, marked the dawn of AI-assisted art. During the 1980s and 1990s, advances in AI technology, particularly the emergence of neural networks and machine learning, opened new avenues for artistic expression. Artists began using these technologies to create more complex and sophisticated artworks. For example, Cohen's AARON was enhanced with machine learning capabilities, leading to more diverse and realistic artworks [322].

The landscape of AI art witnessed a seismic shift with the introduction of deep learning in the past decade. This advanced subset of machine learning, which deploys multilayered neural networks for data analysis and interpretation, has revolutionized a variety of sectors, including the art of AI [323, 324]. Pioneering projects like Google DeepDream and "The Next Rembrandt" exemplify the transformative power of deep learning, facilitating the creation of intricate and lifelike AI art [325, 326]. These projects showcase the ability of deep learning algorithms to interpret and reimagine

visual data in creative and innovative ways. The development of Generative Adversarial Networks (GANs) in 2014 marked another significant milestone in the evolution of AI art [327]. Conceived by Ian Goodfellow, GANs have inaugurated a new epoch in art generation, with their potential most notably demonstrated in the creation of the "Portrait de Edmond de Belamy". This AI-produced artwork, a product of the French art collective Obvious, fetched a staggering $432,000 at a Christie's auction in 2018 [328]. Such developments underscore the growing acceptance and fascination with AI's capabilities within the world of art.

Early GANs faced challenges such as high computational cost and limited output control. The advent of new algorithms and technologies addressed these limitations, leading to a surge in AI art popularity. One such breakthrough was the introduction of CLIP by OpenAI in 2020, which significantly impacted AI art by allowing the generation of art from text prompts [329]. Artists like Katherine Crowson, Mario Klingemann, and Robbie Barrat were instrumental in the popularization and development of these technologies, making AI art accessible to a wider audience [330, 331, 332, 333, 334]. The year 2022 marked the rise of diffusion models, offering a promising alternative to GANs to generate AI art [335]. These models, including Latent Diffusion and Stable Diffusion, provided more stability and diversity in the generated artworks. OpenAI's Dall-e and Stability AI's diffusion model played significant roles in popularizing this approach, helping AI art gain mainstream recognition [336, 337, 338].

As AI's ability to generate art progresses, an important question surfaces: Can AI produce art that is indistinguishable from that created by humans? This critical question forms the central focus of our research. Through a unique experimental setup, this study delves into the indistinguishability of AI-generated art from human-created art, specifically art created by children. Using child-created artwork as the basis for AI-generated pieces, the study aims to examine the depth of AI's creative capabilities and its ability to replicate the spontaneous creativity inherent in art.The paper unfolds as follows. We begin by providing an in-depth background on the

evolution of AI's involvement in art, with a particular focus on generative art. We trace the journey from early algorithmic approaches to the current state of AI-driven creations. Section B.1 outlines the materials and methods used in our study, detailing our unique experimental design. Following this, the results section presents a comprehensive analysis of the experiment's results, with an emphasis on the evaluation of AI-generated art pieces. The discussion section unpacks these findings, interrogating their implications on our understanding of AI's role in art. Finally, we conclude the article by identifying potential avenues for future research and the challenges that lie ahead.

Related Work

The notion of a machine emulating human intellect was first critically proposed by the British mathematician Alan Turing through the famous Turing test [339]. This method was intended to assess the behavior of an AI in a text-based conversation; if its responses could not be distinguished from those of a human, the machine was deemed intelligent.

Since Turing's original proposal, the test has evolved significantly. A plethora of variations have emerged that refine, reinterpret, and expand on Turing's original ideas [340, 341, 342]. One prominent derivative is the Total Turing Test (TTT), which takes into consideration more complex facets of human-like behavior. It not only involves a natural language conversation but also includes other forms of human-like behavior, such as gestures, facial expressions, and even physical movements. The idea is that a machine that can pass the Total Turing Test would be able to seamlessly interact with humans in a wide range of situations and would be virtually indistinguishable from a human being [343]. Moreover, the Turing Test has found prominence in annual competitions such as the Loebner Prize. Here, the machine that most convincingly imitates human-like conversational capabilities is awarded, encouraging developers around the world to design more sophisticated conversational agents [344].

A specific offshoot of the Turing Test is increasingly applicable to the

realm of AI art, the Artistic Turing Test. This test poses the question: can AI-produced artwork be indistinguishable from human-created art? [345, 346]. In a notable instance of this, Elgammal et al. generated art using a Generative Adversarial Network (GAN). The artwork was exhibited and sold alongside human-created art, often unbeknownst to the viewers of its AI origin, thus passing the Artistic Turing Test in a real-world setting [347]. The field of AI literature has also explored the application of the Turing Test. Köbis and Mossink, for example, investigated the ability of AI to generate poetry that could be mistaken for human-created content. Their work further underscores the growth of AI capabilities within creative domains [348, 349]. Recently, work by Li, in which an AI system was tasked with generating creative dance choreography, further expanded the boundaries of the Artistic Turing Test. The AI system was trained using motion capture data from human dancers and produced choreography. This application of the Turing Test underscores the potential of AI in a wide range of creative endeavors [350].

The Turing Test, initially postulated in the context of conversational AI, is gradually becoming a critical yardstick in the AI art domain. This evolution raises intriguing questions about the innovation and creativity capabilities of AI within the creative sphere. Such questions are at the core of this research, which aims to empirically examine the human-like quality of AI-generated artwork.

Materials and Methods

The work presented in this section is currently under review in Computers in Human Behavior: Artificial Humans.

Participants

Participants: Eighty-seven participants were enrolled in the study. Prior to participation, they were provided with all necessary information regarding the test and data handling procedures. No personally identifiable in-

formation was collected; all responses were saved anonymously using a Google Form, ensuring the privacy and anonymity of the subjects involved.

Sampling Procedures: Participants were recruited through online advertisements and word-of-mouth. Inclusion criteria required participants to be over 18 years old and to provide informed consent.

Sample Size: The sample size of 87 participants was chosen to ensure sufficient power for detecting differences in the ability to distinguish between human and AI-generated artworks.

Materials

Primary and Secondary Measures: The primary measure was the ability of participants to distinguish between human and AI-generated artworks. Secondary measures included the analysis of participants' preferences and the qualitative feedback provided.

Quality of Measurements: The quality of measurements was ensured by standardizing the presentation of artworks and randomizing their order. Reliability checks were performed to ensure consistency in the data collection process.

Procedure

Data Collection Methods: Data were collected through a pairwise comparison method. In each pair, the human artifact and the corresponding AI-generated artifact were presented with random placement order to the participants. Participants were asked to identify which of the two images they believed was created by a human.

Research Design: This study used a descriptive and experimental research design. The descriptive component involved analysing the characteristics of the human and AI-generated artworks. The experimental component involved assessing participants' ability to distinguish between the two types of artworks.

Data Processing and Diagnostics: The main statistical characteristics

of the two collections of images were first analysed to understand possible biases in the compositions. Data were then processed to remove any incomplete or inconsistent responses. Outliers were identified and removed based on predefined criteria.

Data Analysis Strategy: Two types of analysis were conducted: a global analysis and a per-picture analysis. The global analysis assessed the overall ability of the participants to distinguish between human and AI-produced artwork. The per-picture analysis provided granular insights on a picture-by-picture basis, identifying specific attributes or patterns influencing participants' decisions. Statistical tests included accuracy, precision, recall, F1-score, and the Area Under the ROC Curve.

AI Art Generation

Model Description: For AI art generation, we used Dall-E 2, an advanced generative model developed by OpenAI. Dall-E 2 builds on the learning capabilities of contrastive models such as CLIP, which are known for their ability to learn robust image representations encapsulating both semantics and style [351]. The model architecture consists of two stages: a prior that generates a CLIP image embedding from a provided text caption, and a decoder that generates an image based on the image embedding.

Generation Process: To generate the images, three steps were followed:

1. Formulating a text prompt based on the description of the human-created artifact. Each prompt contained specific information about the age of the child artist and the painting technique used in the human artifact.

2. Using the Dall-E 2 API to input the prompt.

3. Collecting the image generated based on the given prompt.

AI Classification

Model Description: For the AI classification, we employed an algorithm based on the deep resilient neural architecture ResNet-18 [352]. The algorithm, initially trained on ImageNet [353], was modified so that the final layer of the network had a binary output and was then fine-tuned with our image dataset.

Training and Evaluation: The training configuration consisted of 5 epochs and a batch size of 5. The optimiser employed was Adam with a learning rate of 0.01. Leave-One-Out Cross-Validation was performed to ensure meticulous evaluation, iteratively training the model on each image, thus enhancing performance robustness. Performance metrics computed included accuracy, precision, recall, F1-score, and the Area Under the ROC Curve.

Results

Figure B.1 shows the 17 artworks produced by the child using oil paint. The collection of artworks displays a wide diversity of subjects, vividness of colours, and a range of complexity. Human creativity shines through, each piece offering a unique glimpse into imagination and the way the world is perceived. Each piece presents a unique subject matter that depicts a myriad of scenes from everyday life, fantastical scenes, and vibrant landscapes. The breadth of themes suggests an expansive and explorative creative mind. The child is not limited to the mundane or real, delving fearlessly into the realm of the imagined, challenging the viewers' perception and expectations. The use of colours is strikingly vivid, with bold hues setting a strong visual impact. The colour palette ranges from bright, saturated colours to more subdued, earthy tones, creating a vibrant tapestry of artwork. Each chosen colour, whether purposefully or instinctively, adds a new dimension to the artwork, encapsulating the emotional state, mood, or whimsy at the time of creation. The complexity of the artwork is equally impressive. There are simple designs with clear and straightforward narratives, while

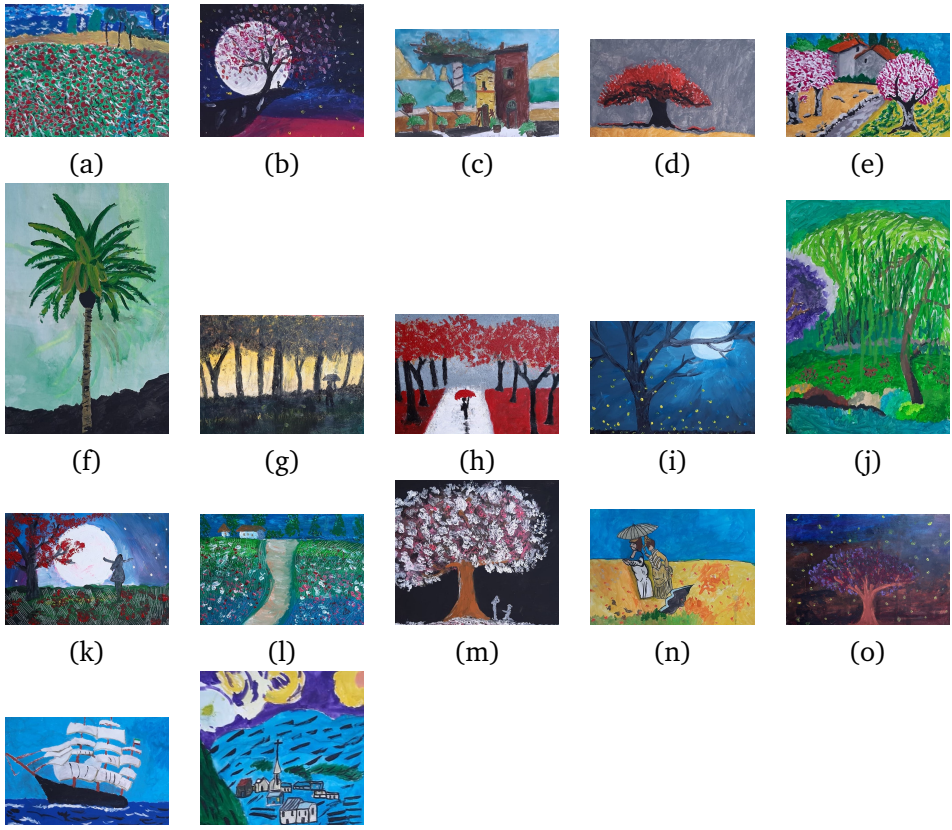


Figure B.1. Matrix of child produced artworks.

others have layers of complexity that invite deeper introspection. This combination of simplicity and complexity not only reflects the child developing artistic skills but also their willingness to experiment and push their creative boundaries. Overall, these 17 artworks speak volumes about the child inherent creativity and growing technical ability. They remind us of the limitless potential of a young mind given the freedom to express itself.

Prompt Engineering and Image Generation

The prompts for image generation were designed based on the guidelines of [354]. Each prompt highlights key aspects of the corresponding

work of art, being devised to encapsulate not only the subject of the painting but also its main elements, predominant colours when significant, the style of art, and the notable detail that it has been drawn by a child. The first aspect addressed in the prompt is the subject of the painting. This could be anything from a simple object to a complex scene, an abstract design, or even a fantastical creature. The goal is to provide a concise but accurate description that can guide the AI model. The main elements of the artwork are then detailed; this includes the notable features and components that make up the artwork, and could range from the individual characters in the scene to the intricate patterns in an abstract design. When applicable, the main colours used in the artwork are mentioned. This is especially crucial for pieces where colour plays a significant role in conveying the mood or theme. The paint technique, which is always oil paint, is also indicated. This provides further context for the AI model, helping it to generate artwork that aligns with the same technique. Finally, each prompt emphasizes that the artwork is the work of a child. This is a crucial detail, as it sets the context for the child-like creativity and unbridled imagination that the AI model should strive to emulate. This rigorous and methodical process of prompt creation helps to ensure that the AI model is equipped with a comprehensive and nuanced understanding of each artwork, thereby increasing the chances of generating comparable AI-created artworks.

The complete list of prompts is shown in Table B.1, while the images generated by DALL-E 2 are reported in Figure B.2. It is worth observing that, by sending the same prompt to DALL-E, the generated image(s) may vary as a result of generation-noise. As a working method, we decided to take for each prompt the first creation, thus avoiding to introduce a bias related to an arbitrary choice made by a human user on a set of generated images.

AI-Based Generation Consistency

The primary objective of using the DALL-E 2 model in our study is not to precisely duplicate the child's artwork, but to construct an image

Table B.1. Image labels and corresponding prompts.

Image label	Prompt
(a)	Portrait of a poppy field, blue sky and flock of white birds as a child would draw it using oil paint
(b)	Portrait of a night sky, a child under a flowering cherry tree above a cliff as a child would draw it using oil paint
(c)	Portrait of a pair of colourful houses, with a courtyard with a red apple tree and many pots filled with green plants as drawn by a child in oil paints
(d)	Portrait of a lone tree on a yellow ground covered with red leaves, behind the sky is gray as drawn by a child with oil paints
(e)	Portrait of a country driveway with two houses on either side and peach blossom trees as drawn by a child in oil paints
(f)	Portrait of a palm tree on a rocky background and under blue sky as drawn by a child with oil paints
(g)	Portrait of a man with umbrella against the sunset in the midst of trees as drawn by a child in oil paints
(h)	Portrait of a man with red umbrella walking among trees with red leaves planted on red earth as drawn by a child with oil paints
(i)	Portrait of a leafless tree under the moon at night with fireflies as drawn by a child with oil paints
(j)	Portrait of a willow tree on a lake and a tree with purple flowers as drawn by a child with oil paints
(k)	Portrait of a dancer on a poppy meadow under the moon at night as drawn by a child in oil paints
(l)	Portrait of a country road to a group of houses and cypress trees as drawn by a child with oil paints
(m)	Portrait of a love proposal under a flowering tree at night as drawn by a child in oil paints
(n)	Portrait of two nineteenth-century women walking along a beach under the sun as drawn by a child in oil paints
(o)	Portrait of red tree at night with fireflies as drawn by a child with oil paints
(p)	Portrait of sailing ship with two masts in the middle of the sea as drawn by a child with oil paints
(q)	Portrait of village with church from above a hill as drawn by a child in oil paints

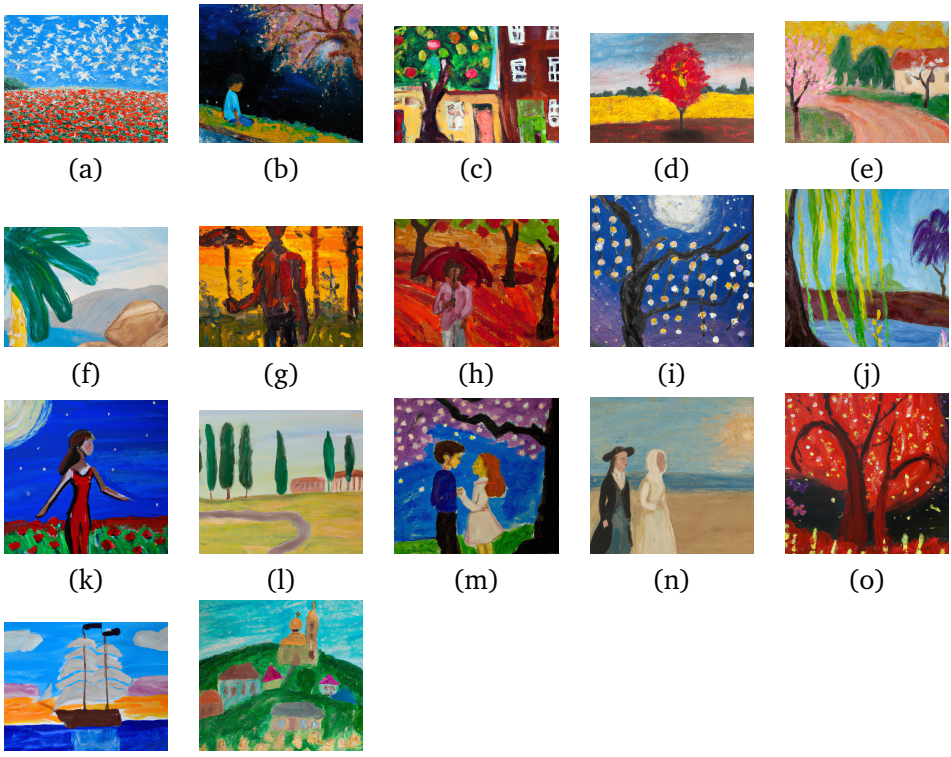


Figure B.2. Matrix of DALL-E 2 produced artworks.

that exhibits coherence and plausibility given the provided prompts. These prompts were designed to serve as guiding principles for the AI model, not as a rigid blueprint. The effectiveness of AI image generation should not be gauged on its ability to mirror the original artwork in an exact pixel-to-pixel manner. Instead, we sought a certain degree of likeness in terms of context, style, technique, and evoked emotions. The expectation is for the AI model to create artwork that shares thematic, stylistic, and emotional resonance with the original child art while maintaining its unique AI-influenced interpretation. The inherent variability of DALL-E 2 facilitates this creative divergence. The model, although guided by the prompts, is capable of interpreting the input in a myriad of ways, which aligns with the unpredictable, yet inspiring nature of artistic creation. By comparing the human and AI-

generated artwork, we observed that DALL-E 2 was able to generate pieces that, while not identical to the originals, mirrored the child's art in interesting and often surprisingly creative ways. It embodies the critical elements described in the prompts, thereby demonstrating a certain consistency in its generative capability, whilst adding its unique nuances.

Test on humans

The experiment on human's results can be seen in Figure B.2. Results

Table B.2. Human test results.

Accuracy	Precision	Recall	F1-Score
49.76%	51.79%	52.45%	52.12%

show that human ability to discern human generated images from AI generated images is very low: accuracy is close to 50%, indicating that they choose the image almost randomly.

Test on AI

Classification results can be seen in table B.3. Results demonstrate a

Table B.3. AI classification results.

Metric	Accuracy	Precision	Recall	F1-Score
Global Metrics	97.06%	97.22%	97.06%	97.06%
AUC-ROC	99.65%			

high level of accuracy, with an overall accuracy rate of 97.06%. Precision and recall metrics are equally impressive, indicating that the model is adept at correctly identifying both human and Ai generated instances. The F1-score, a harmonic mean of precision and recall, further validates the model's balanced performance. The Area Under the ROC Curve (AUC-ROC) attaining 99.65% underscores the model's exceptional ability to discriminate between the two classes.

Discussion

The results of this study highlight the significant advancements in AI's ability to generate art that closely mimics human creativity, specifically that of a child. The experimental design and subsequent analysis provide valuable insights into the nuanced capabilities and limitations of AI in the creative domain.

The human participants' performance, with an accuracy close to 50%, indicates their difficulty in distinguishing between AI-generated and child-created artworks, consistently with the findings in [355, 356]. This result is particularly compelling as it suggests that AI-generated art has reached a level of sophistication and authenticity that can deceive even the human eye. The precision, recall, and F1-score metrics further underscore this point, as the values hover around the mid-50% range, reflecting the random nature of the participants' choices. These findings align with previous research on the Artistic Turing Test, where AI-generated art successfully passed as human-created in several instances [347, 346].

In stark contrast to human participants, the AI classifier exhibited remarkable proficiency in distinguishing between human and AI-generated art. With an overall accuracy of 97.06%, the AI model demonstrated its ability to effectively identify subtle differences between the two types of artworks. The high precision and recall values indicate that the model is adept at correctly identifying both true positives (correctly identifying AI-generated art) and true negatives (correctly identifying human-created art). The F1-score of 97.06% further validates the model's balanced performance, while the AUC-ROC value of 99.65% highlights its exceptional discriminative power. These results suggest that AI, when trained appropriately, can develop a keen sense of artistic style and origin, surpassing human capabilities in specific tasks [352, 353].

The findings of this study have several profound implications for the future of AI in the creative industries. First, the ability of AI to generate art that is nearly indistinguishable from human-created art opens new possibilities for collaborative art creation, where AI can act as a co-creator rather

than merely a tool. This shift could lead to novel forms of artistic expression and innovation [319, 318]. Second, the study underscores the importance of developing robust methods for detecting AI-generated content. As AI-generated art becomes more prevalent, ensuring the authenticity and provenance of artworks will be crucial. This necessitates the development of advanced AI detection algorithms, similar to the classifier used in this study, to maintain integrity in the art market and other creative fields [329, 335]. Finally, this research highlights the potential of AI to enhance our understanding of human creativity. By analysing the characteristics that AI uses to distinguish between human and AI-generated art, we can gain deeper insights into the elements that define artistic creativity and expression. This could lead to new theories and models of creativity that integrate both human and machine perspectives [345, 350].

Conclusion

This study provides compelling evidence that AI has achieved a level of artistic sophistication that allows it to produce art indistinguishable from human-created works, particularly those by children. The difficulty faced by human participants in identifying AI-generated art, contrasted with the high accuracy of the AI classifier, underscores the advancements in AI capabilities. These findings open new horizons for the integration of AI in the creative industries, presenting both opportunities and challenges that warrant further exploration.

Acknowledgment

We would like to express our deepest gratitude to the young artist whose creative talents provided the foundation for this study. Your imaginative and inspiring artworks were integral to our research. We also extend our heartfelt thanks to the artist's mother for her support and assistance throughout this project. Your encouragement and cooperation were invaluable in making this study possible. Thank you both for your significant contributions.

B.2 Mathematics As A Crossroads Between Visual Arts And Music

The link between science and art emerges throughout history as a manifestation of abstract and symbolic thinking by homo sapiens in an attempt to provide a rational representation of the world around him. The same spirit of observation and curiosity that unites artists and scientists leads to the intuition that underlies creativity. In science, intuition is validated using Galileo's experimental or hypothetical-deductive method, which follows a mathematical logic, linked to principles of harmony and proportion. This concept, known since the origin of abstract thought, when our ancestors started to paint cave walls and carve figures in stone, was later elaborated by the Greeks and taken up by the Renaissance. The sculptor Polyclitus, who lived in Athens in the second half of the 5th century BC, collected in a treatise, the Canon, the basic rules of artistic creation capable of ensuring rhythm, proportion and harmony. In music, too, it has been known since the time of Pythagoras that harmony, which is based on numerical ratios identifying the intervals between notes, is, as in other forms of art, the basis of beauty that arouses pleasure. For example, Pythagoras discovered, by playing strings of different lengths together, that the interval of a perfect fifth (ratio of 2:3 in length) elicits a pleasurable experience of consonance. More generally, Pythagoras argued that intervals that can be expressed as ratios of small integers are perceived as consonant and pleasant, while other intervals are perceived as dissonant and unpleasant. This is compatible with a modern theory that considers these ratios as easier to process by the ear-brain system.

Since antiquity, man has assumed that the laws governing musical phenomena and its very proportions can be found in the harmony of the cosmos. Pythagoras and his followers thought that the relationships that generate harmony in music also occurred in the celestial sphere, in the distances between celestial bodies and the centre of the world. In the 17th century,

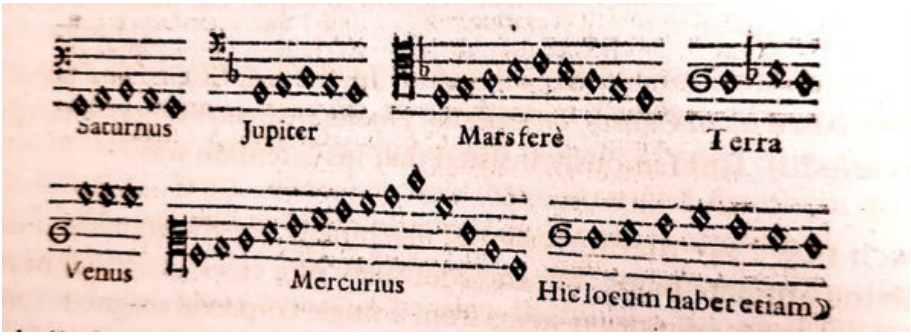


Figure B.3. Representation of the notes played by each planet in its orbit around the sun, from Kepler's Harmonices mundi.

Kepler, in the fifth book of his work Harmonices Mundi, expounded his theory on the "music of the celestial spheres", accompanied by numerical tables and musical scores, with the "notes" played by each planet. For Kepler, it is evident that, if planets are moving on a circular orbit at constant speed, they would only produce monochords. If, on the other hand, they move on an elliptical orbit with variable speed, each planet generates a melody! Only Venus generates a monochord as its orbit is almost circular (Figure B.3).

Interestingly, at a certain point in the treatise, Kepler interrupts his speculations on the music of the planets to introduce what is now known as Kepler's third law of planetary motion. In this regard, great scientists of the past (Kepler, Darwin) did not hesitate to include aesthetic considerations in their scientific works. This is an attitude that has been completely lost today, perhaps due to the mechanistic tendency of modern science.

Thus, the microcosm (the world in which we live and, in particular, man) and macrocosm (the entire universe) appear to be linked by common mathematical and aesthetic rules. Medieval philosophers thought this as a logical consequence of man and the universe having been created by the same Creator. Kepler, when investigating the laws of planetary motion, wanted to discover nothing less than the thoughts that God himself had when creating the universe. The idea of the profound relationships existing between the micro- and macrocosm, which derives from Platonic

philosophy on the perfection of the world of ideas, as opposed to the imperfection and contingency of the concrete world, was particularly developed in the medieval period by Neo-Platonic philosophers, and some historians of science argue that it led to the decline of the sciences in the medieval period because, when interpreted in an extreme form, it pushes us to seek at all costs relationships between the micro- and macrocosm that are fanciful and arbitrary. However, as pointed out above, it is undeniable that Neo-Platonism provided the inspiration for Kepler for his experimental research on the motion of the planets and the discovery of their laws.

In the Sententia Libri de sensu et sensato (1268), St. Thomas Aquinas wrote: "The senses delight in things that have the right proportions". Following the translation of the manuscripts of Euclid, Ptolemy and Archimedes, Euclidean geometry was considered, as suggested in the Opus Maius (1267) of Roger Bacon, a Franciscan friar and philosopher who lived in the 13th century, to be the point of convergence of objective truth not only in figurative arts but also in sciences. These theories certainly contributed to the development of perspective as "the creation of concrete bodies perceptible to our eyes", which Giotto certainly made use of in his frescoes on the life of St. Francis in the Basilica of Assisi. Furthermore, geometry and its numerical ratios had a crucial importance in sculpture, considered the means of artistic expression that in the Renaissance led to the consideration of "man as the measure of all things". This led to the concept of beauty (in nature, in a painting, in a statue or in an architectural element) as harmony between the parts realised through the so-called golden ratio, developed in particular by Luca Pacioli, a Franciscan friar and mathematician who lived at the time of Leonardo, of whom he was a friend, who in De Divina Proportione (1509) theorized the application to all arts of the golden ratio of the radius, which is the proportional average between the entire segment and the remaining part. The algebraic expression of the golden ratio is equal to:

$$(a + b) : b = b : a \tag{B.1}$$

where a represents the smaller side of a rectangle and b the longer side.

This number is denoted by Φ because it is not possible to write its exact value as it is an irrational number, slightly greater than 1 but consisting of an infinite number of decimal places.

$$\Phi = \frac{1+\sqrt{5}}{2} \tag{B.2}$$

Such proportion was considered divine in that it reflects divine perfection and harmony and allows one to understand the world through the law according to which it was created. Pacioli advocated the application of Euclidean geometry to all forms of activity, which he even extended to the letters of the alphabet (Figure B.4).

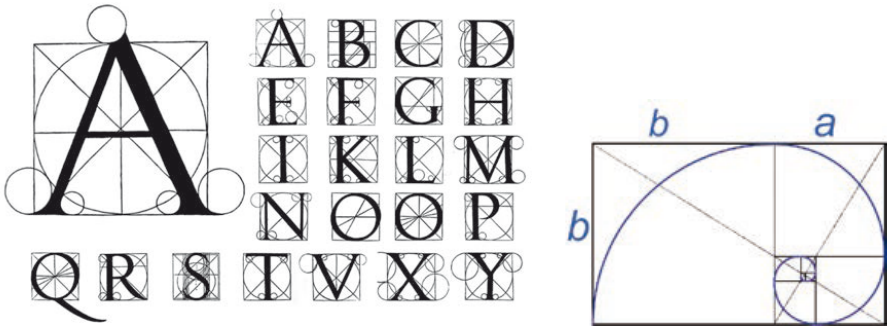


Figure B.4. Renaissance "capital" letters, from De divina proportione by Luca Pacioli.

This theory was to be put into practice by artists such as Leon Battista Alberti, Piero della Francesca and Leonardo himself. In particular, Leonardo not only sought to determine how the golden ratio is capable of determining ideal beauty in the forms of the human body, but was firmly convinced that natural phenomena including the motion of water and the flight of birds were also subject to the laws of geometry, further fostering that privileged relationship linking mathematics to art. This confirmed Plato's theory that aesthetic canons are present in the very object of perception and are independent of social and cultural factors and are therefore invariable over time. In recent times, the architect Le Corbusier (1887- 1965) used the golden ratio in the Modulor system he developed, to calculate the size of

ideal human body parts, considering different postures, taking a man of height 182.9 cm as a reference (Figure B.5).

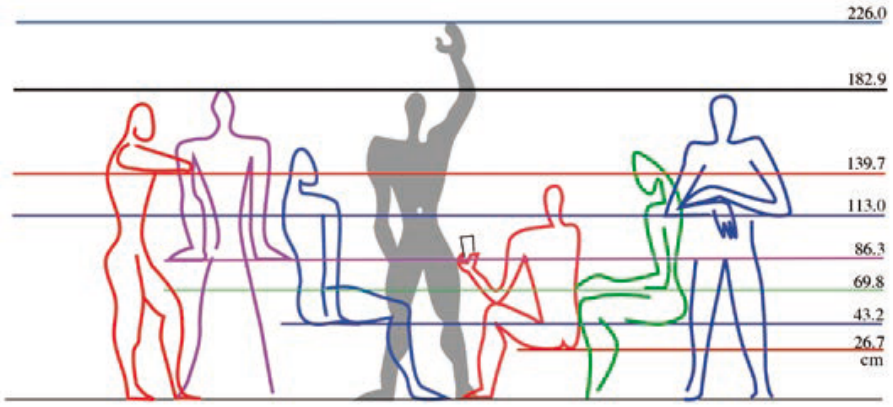


Figure B.5. Modulor system. The dimensions that are in exact golden ratio to the human height of 182.9 cm are 113.0 cm; 69.8 cm; 43.2 cm; 26.7 cm (Skyamal - Own work).

Le Corbusier later used the Modulor system in the façade of several buildings (Figure B.6).

The same numerical relationships underlying the harmony between the parts can be found in nature, for example in the shell of Nautilus pompilius (Figure B.7), in the growth of plants, in a cyclone, in the galaxy, which reflect to some extent the mathematical concepts underlying the logarithmic spiral and the Fibonacci sequence. The latter consists of a succession of numbers, each of which is the sum of the previous two: 0, 1, 2, 3, 5, 8, 13... The relationship between the golden number and the Fibonacci sequence, which remained unknown even to Luca Pacioli, was discovered by Kepler in 1611, who realised that the ratio between two consecutive numbers in the Fibonacci sequence gradually approximate the golden number Φ.

In accordance with Plato's view, the demonstration that beauty has its own objective biological basis, independent of experience, was provided a few years ago by Cinzia Di Dio and Giacomo Rizzolatti of the University of Parma in collaboration with Emiliano Macaluso of the Fondazione St. Lucia in Rome (Di Dio et al., 2007). Using proportion as an independent vari-

Figure B.6. RLe Corbusier, Unité d' habitation, Marseille (Gil Singer/Alamy Stock Photo).

Figure B.7. Nautilus pompilius shell (Chris 73/Wikimedia Commons).

able and neuroimaging techniques such as Functional Magnetic Resonance Imaging (fMRI), these authors subjected 18-20 year-old fellows, who were not art experts, to two types of stimuli: one consisting of the image of the statue of the Doryphoros (spear bearer) by Polyclitus, universally consid-

ered a masterpiece of harmony between the different parts of the body, and the other, a modified version of the same image, in which the golden proportion between certain parts of the body (back and limbs) was altered. The subjects also had to indicate whether they considered the image beautiful or ugly. Whereas in the latter case only the brain areas associated with visual perception were activated, in the former, in addition to the visual areas, areas involved in emotions and pleasure such as the insula, amygdala and hippocampus were activated. Furthermore, the canonical images were evaluated positively by the observers while the modified images were evaluated negatively. It is therefore clear from these results that the objective parameters intrinsic to works of art are able to evoke a specific neural pattern that determines the sense of beauty in the observer. The key to the change in the perception of a sculpture from "ugly" to "beautiful" seems to be linked to the joint activation of cortical neuronal populations that respond to specific features present in works of art and of neurons located in emotional control centres. It is therefore plausible that reading a text whose letters follow the "golden" canons described by Pacioli also leads to the activation not only of the visual and language centres, but also of those of pleasure with the consequent enhancement of the mental representation of concepts and images generated by reading. The aesthetic parameters generated by the mathematical ratio underlying Φ, associated with standards of beauty, found in various natural and biological systems, suggest that these have played a fundamental role in the course of evolution. In particular, the golden ratio has led to an optimisation of structure, as in phyllotaxis (from the Greek: phyllon = leaf and taxis = order) of plants in which the arrangement of leaves on a stem follows a circular component that traces an imaginary helix around the stem. Starting from any leaf, after one, two, three or five turns, depending on the different plants, one always finds a leaf aligned with the first. As we move upwards, the number of turns made to find the leaf aligned with the first one corresponds to the Fibonacci succession, which can be approximated to the golden number 1.618... This succession favors optimal exposure of the leaves to air, rain and light for

the benefit of photosynthesis.

A recent study of the architecture of the skull in humans has shown how, over millennia, it has followed an elegant harmonization between structure and function expressed in the golden ratio especially in the so-called neurocranium, the portion that protects the brain and the four sense organs, called the calvarium. As highlighted in Figure B.8, the ratio between the length of the nasion-inion and bregma-inion arch is equal to Φ and this coincides with the ratio between the bregma-inion and nasion-bregma arch (Tamargo and Pindrik, 2019). As in humans also in anthropoid apes (orangutans, gorillas, chimpanzees), the nasion-bregma frontal arch is more developed as it reflects the growth of the frontal lobes. In other mammals, a convergence towards Φ would correlate with increasing species complexity. Therefore, the conservation of Φ throughout evolution would suggest that it is somehow related to the survival of the species.

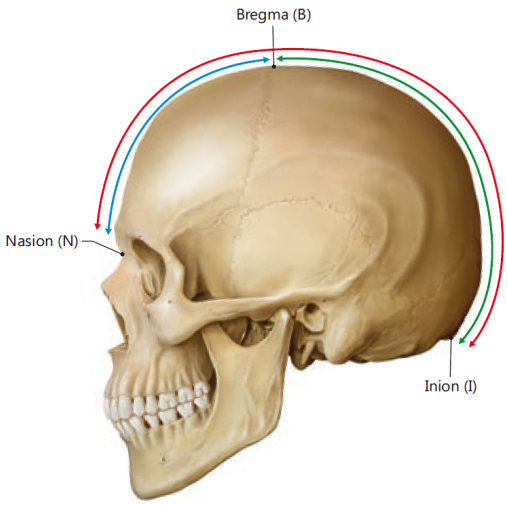


Figure B.8. In the human skull, the geometric ratio between parieto-occipital arch (between bregma and inion, BI) and frontal arch (between nasion and bregma, NB) coincides with the relationship between nasion-inion (NI) and bregma-inion (BI). Both correspond to the golden ratio of 1.618... (modified from Tamargo RJ et al. 2019).

According to the commonly accepted view of natural selection, for a

character to evolve it must be adaptive, i.e. it must increase the individual's probability of survival. Therefore, the importance of the aesthetic factor, present not only in humans but also in many animal species, in the evolution of species is difficult to explain. In The Origin of Species Darwin put it this way: "How the sense of beauty in its simplest form – that is, the reception of a particular pleasure derived from certain colors, shapes and sounds – first developed in man and the lower animals, is a question that is by no means clear". Nonetheless, in The Descent of man, Darwin traced the sense of beauty back to the courtship between the two sexes, regulated by sexual selection for the purpose of reproduction. To seduce females, males make use of aesthetic means such as sounds, colors and shapes. According to Darwin, the conspicuous beauty of the upper part of the wings of male butterflies is the result of the females' preference for the more beautiful males. Over time, the choice of females increased the likelihood of beautiful wings in their offspring, thus contributing to the large number of colors and shapes in butterfly wings that we can still appreciate today.

Therefore, the sense of beauty constitutes an adaptation useful for our survival and directly shaped by natural selection. This concept has recently been taken up by Richard Prum who, in the Evolution of Beauty (2020), revisits the theory of evolution showing how it is based not only on the principle of adaptation through natural selection but also on sexual attraction in which the shape of the object of desire plays a significant role. A species would express a preference for a certain character judged to be beautiful, and based on that preference one of the two sexes, usually the female, would make the choice of partner.

Ranging between evolutionary biology, philosophy and sociology, Prum rewrites the theory of evolution, redeeming the role of beauty and desire, and offers us a new, fascinating natural history centreed on female arbitrariness and the sense of beauty as opposed to the law of struggle and the domination of the strongest. However, it should be noted that such an approach, if unduly transferred to social policies, opens the door to eugenics and selection on the basis of physical appearance, examples of which we

unfortunately also have in recent times.

Turning now to the consideration of the relationships between figurative arts and music, one can point to strong analogies between the two fields: for example, many authors have pointed out the relationships between the various musical forms (sonata, fugue, symphony etc.) and architectural forms, so much so that it is common to speak of "architecture" or "supporting lines" of a musical piece. Conversely, it is also common to speak of "rhythm" in architecture, understood as the repetition of the same motif at regular intervals.

If we consider more closely the role of numerical ratios in music, we see that here too there is a clear relationship between numerical ratios and aesthetic perception. We can start with the simple difference, obvious to all, between a musical note and a noise. Both acoustic phenomena lie within the range of audible frequencies, but whereas a note – more precisely a tone – can be assigned a defined pitch within a musical scale (e.g., A3, E5 etc.), noise cannot be assigned a pitch. What is the difference between tone and noise from a physical-mathematical point of view? The difference is demonstrated in the Figure B.9. If we subject both acoustic phenomena to a frequency spectrum analysis (Fourier analysis), we can recognise in the tone a fundamental frequency, which defines its pitch, and other frequencies, called harmonics, which are in precise geometric ratios (2:1, 5:4 etc.) with the fundamental frequency, whereas in noise we can recognise neither a fundamental frequency nor harmonics (Figure B.9).

Mathematics is also fundamental to the division of tonal intervals. In the diatonic scale of Western music, based on the octave ratio 2:1, two tones whose fundamental frequencies are in the ratio of 2:1, 4:1 etc., although perceived at different pitches, are perceived as equivalent, and are therefore named in the same way (C, D etc.) despite belonging to different octaves. In general, two notes a and b belong to the same equivalence class if their frequency ratio $a/b = 2n$. We recognise the same sequence of notes played at different octaves as identical. Each successive octave covers twice the frequency range of the one below it. Octave intervals are the same if they

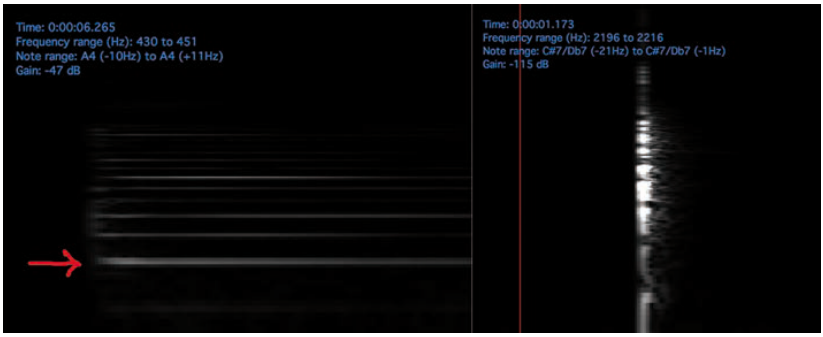


Figure B.9. Frequency analysis of a musical tone (A4 from a piano, left) and a noise (wooden hammer banging on a table, right). Vertical axis = frequency. Horizontal axis = time. A fundamental frequency (arrow) and harmonics at higher frequencies can be recognised on the left. Image generated with WavePad software.

are on a logarithmic scale. Modern instruments are generally tuned to A 440 Hz, but other tunings are possible (e.g., 415 Hz for baroque instruments).

A musical scale is a regular sequence of successive pitches that repeat at cyclic intervals. The subject of the division of musical intervals has been a widely debated topic since ancient Greece. It is a fundamental issue for musical composition, as it is functional to transpose any musical interval into any key without altering the relationship between notes. Such a subdivision allows the transposition of any note, or chord, or melody, from a given scale (e.g., C major) to another scale (e.g., G major) while maintaining the intervals between the notes and thus also the harmonic and melodic characteristics unchanged, so that a melody performed in the C or G scale appears musically identical, at least to most people (with the exception of those with absolute pitch). Simplifying greatly, two completely different principles have historically been considered for generating musical scales. The first principle is called just temperament and it is based on the introduction of two more intervals that are consonant and therefore pleasing to the human ear, in addition to the unison and octave interval, namely the intervals of exact fifth (frequency ratio 3/2) and exact fourth (frequency ratio 4/3). The use of the ratios 3/2 and 4/3 derives from the Pythagorean

principle that only ratios of small numbers generate consonances. By using a cyclic principle of alternating ascending fifths and descending fourths, all semitones covering the octave interval (Pythagorean chromatic scale, see Figure B.10) can be generated. In this way, the frequency ratio between two tones separated by an integer interval, e.g. between Eb and F, or F and G is 3/2*3/4 = 32 /23 = 9/8, while the semitone, e.g. between Eb and E, is 3/2*3/4*3/4*3/4 = 37 /211 = 2187/2048.

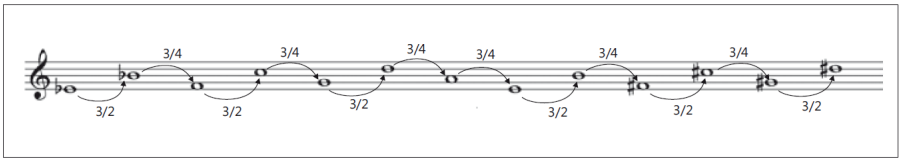


Figure B.10. Pythagorean chromatic scale.

This process can be expressed on a circle, called the circle of fifths. The scale shown above can be represented as follows (Figure B.11).

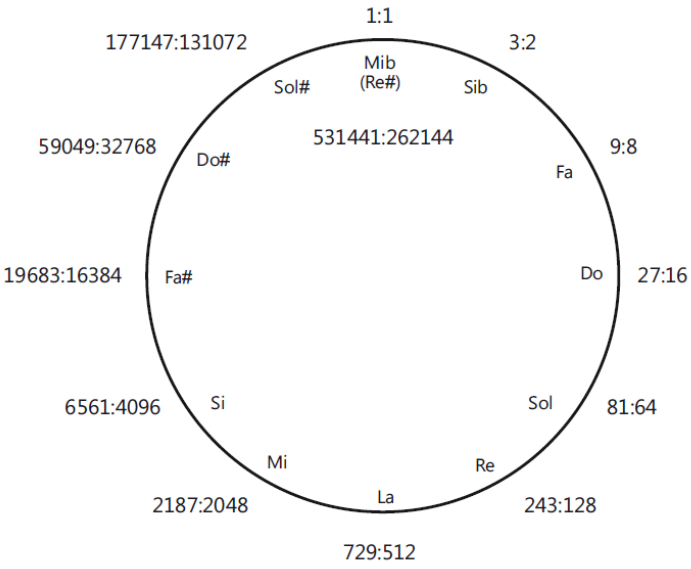


Figure B.11. Circle of fifths.

In essence, in the Pythagorean chromatic scale, semitones are obtained by multiplying the fundamental frequency by rational numbers. The problem with this type of scale is that, at the end of the cycle of fifths and fourths, after twelve semitones, we do not return to the same starting note (E♭ in the example) but to a note that is slightly higher (D♯ in the example). This stems from the fact that the ratio between the starting note and the ending note is

$$\frac{3^{12}}{2^{18}} = \frac{531441}{262144} \approx 2.027,$$

which is not exactly 2, but slightly greater. The difference is called the *Pythagorean comma*. The fundamental reason for this discrepancy is that no natural number can simultaneously be an integer power of 3 and an integer power of 2.

To overcome this problem, the so-called equal temperament was introduced for the first time by Simon Stevin (1548-1620) and by Vincenzo Galilei (1520- 1591), father of Galileo, and is still in use today in western music. Here, the octave interval is subdivided into twelve identical subintervals, called semitones. The value of a semitone is obtained by considering the scale as a geometric progression

$$\alpha_n = \alpha_0 \cdot q^n,$$

in which the octave ratio is known $\alpha_{12} = 2\alpha_0$. Consequently, one can calculate the ratio

$$q = \sqrt[12]{2} \approx 1.059\ldots$$

corresponding to the value sought. On a logarithmic scale, the intervals separating adjacent semitones are equal, which is in agreement with the general Weber-Fechner law of perception. Note that the semitone in equal temperament has a slightly lower value than in the Pythagorean scale (2187/2048 0 1.067...).

Since the definition of a semitone contains $\sqrt{2}$, the ratio of frequencies is an irrational number. The octave is related to the tonic by 212/12,

i.e., 2:1, the right fourth (5 semitones above the tonic) in 25/12 ratio, the right fifth (7 semitones above the tonic) in 27/12 ratio. The Well-Tempered Clavier is a series of preludes and fugues composed by Johann Sebastian Bach (1685-1750) to demonstrate the harmonic possibilities of the tempered scale, although the type of temperament used by Bach is still debated today.

Up to this point, we have treated consonance and dissonance between two sounds as depending solely on the sharing of harmonics physically present in the sounds. The more harmonics shared, the greater the degree of consonance and pleasantness of a chord. In reality, things are more complicated: the ear-brain system does not function according to simple physical-mathematical rules; an example of this is the phenomenon of the "missing fundamental frequency": if we eliminate the fundamental frequency from a musical tone, we continue to perceive the tone at the same pitch! So, it is clear that the ear uses the entire spectrum of frequencies present to define the pitch of a tone. Furthermore, when we hear two notes simultaneously, we hear a third sound, not present in the original sounds, also known as the combination sound. The first to describe this phenomenon was the 18th century violinist Giuseppe Tartini (Caselli et al., 2018). Where does this combination sound come from? Research in recent decades shows that the cochlea does not simply behave as a passive receiver, but it is able to generate sounds by means of positive feedback from the outer hair cells: these cochlear-derived sounds are the origin of otoacoustic emissions.

Is it possible to imagine musical scales that are based on a fundamental ratio different from the octave ratio? The answer is yes: a well-known example is the Bohlen- Pierce scale, which is based on a fundamental ratio of 3:1 (one octave above a perfect fifth). In this type of scale, each note is unique and is never repeated.

Is it possible to imagine a musical scale based on the golden ratio? If one were to construct a new scale based on the fundamental ratio with Φ:1, and keeping twelve semitones in the scale, they would obtain a new value

of

$$q = \sqrt[12]{\Phi} = \sqrt[12]{\frac{1+\sqrt{5}}{2}} \approx 1.040916... \tag{B.3}$$

This value poses a problem for consonance: in the tones generated by all non-electronic musical instruments and the human voice, there are harmonics, which respect diatonic ratios. Since consonance in an interval between two notes is given by the presence of one note in the harmonics of the other, it is clear that changing the ratios breaks traditional consonances. Synthesizing a tone and summing its harmonics according to the new ratios is expected to re-establish consonant intervals.

Similarly, to what was done by Houtsma in the Auditory demonstrations CD, the authors of the present contribution simulated a scale respecting these new ratios. The simulation was created according to the following criteria: Each tone is composed of the sum of sine waves with a frequency equal to the fundamental frequency and its harmonics. Each sound contains up to its own seventh harmonic.

- The amplitude of each harmonic is inversely proportional to the harmonic number.
- The amplitude trend over time is exponentially decreasing.

We can summarise the following points in the formula:

$$s = \sum_{i=1}^{7} \frac{\cos(2\pi f \cdot \mathrm{coef}_i \cdot t)}{i} \cdot e^t$$

In which:

$$\mathrm{coef} = \left(1, st^{12}, st^{17} \cdot \mathrm{cent}^2, st^{24}, \frac{st^{28}}{\mathrm{cent}^{14}}, \frac{st^{31} \cdot \mathrm{cent}^2}{\mathrm{cent}^{31}}, \frac{st^{34}}{\mathrm{cent}^{31}}\right),$$

$$st = \sqrt[12]{\Phi}, \quad \text{and} \quad \mathrm{cent} = st/100.$$

Using this sound library, Colonel Bogey March and Bach's Chorale No 1 were simulated.

These brief outlines of music theory led us to consider that geometric ratios also play a fundamental role in music, although in Western classical music they do not correspond to Φ.

In conclusion, although Φ is a privileged relationship as far as the canons of harmony in the figurative arts are con-cerned, the canons of harmony in music follow rules that are somewhat different. It therefore seems appropriate to recall the quote from St. Thomas Aquinas quoted at the beginning: "The senses delight in things that have the right proportions". What the "right proportions" are, depends on the sensory modality and what the artist intends to express.

B.2.1 ACKNOWLEDGEMENTS

The Authors are particularly grateful to Professor Luca Vollero for his assistance in simulating the musical scale based on the golden ratio and to Professor Emanuele Stracchi for his comments and suggestions.

BIBLIOGRAPHY

[1] Chonyacha Suebsin and Nathasit Gerdsri. “Key factors driving the success of technology adoption: Case examples of ERP adoption”. In: *PICMET'09-2009 Portland International Conference on Management of Engineering & Technology*. IEEE. 2009, pp. 2638–2643.

[2] Surabhi Singh, Shiwangi Singh, Mayur Chikhale, and Sanjay Dhir. “Critical success factors for emerging technology adoption, strategic flexibility, and competitiveness: An evidence-based total interpretive structural modeling approach (TISM-E)”. In: *Global Journal of Flexible Systems Management* 25.3 (2024), pp. 601–628.

[3] Sharon Manship, Eleni Hatzidimitriadou, Julia Moore, Maria Stein, Debra Towse, and Raymond Smith. “The experiences and perceptions of health-care professionals regarding assistive technology training: a systematic review”. In: *Assistive Technology* 36.2 (2024), pp. 123–146.

[4] Marina Liselotte Fotteler, Viktoria Mühlbauer, Simone Brefka, Sarah Mayer, Brigitte Kohn, Felix Holl, Walter Swoboda, Petra Gaugisch, Beate Risch, Michael Denkinger, *et al.* “The effectiveness of assistive technologies for older adults and the influence of frailty: systematic literature review of randomized controlled trials”. In: *JMIR aging* 5.2 (2022), e31916.

[5] Deepshikha Yadav, Surinder P Singh, and PK Dubey. “Glucose Monitoring Techniques and Their Calibration”. In: *Handbook of Metrology and Applications*. Springer, 2023, pp. 1–23.

[6] Árni Kristjánsson, Alin Moldoveanu, Ómar I Jóhannesson, Oana Balan, Simone Spagnol, Vigdís Vala Valgeirsdóttir, and Rúnar Unnthorsson. “Designing sensory-substitution devices: Principles, pit-

falls and potential 1". In: *Restorative neurology and neuroscience* 34.5 (2016), pp. 769–787.

[7] Jan Auernhammer. "Human-centered AI: The role of Human-centered Design Research in the development of AI". In: (2020).

[8] Irene Göttgens and Sabine Oertelt-Prigione. "The application of human-centered design approaches in health research and innovation: a narrative review of current practices". In: *JMIR mHealth and uHealth* 9.12 (2021), e28102.

[9] Catherine Audrin and Bertrand Audrin. "More than just emotional intelligence online: introducing "digital emotional intelligence"". In: *Frontiers in Psychology* 14 (2023), p. 1154355.

[10] Emad A Abu-Shanab and Amro Abu Shanab. "The influence of emotional intelligence on technology adoption and decision-making process". In: *International Journal of Applied Decision Sciences* 15.5 (2022), pp. 604–622.

[11] Charles Spence. "Crossmodal correspondences: A tutorial review". In: *Attention, Perception, & Psychophysics* 73 (2011), pp. 971–995.

[12] Ella Striem-Amit, Laurent Cohen, Stanislas Dehaene, and Amir Amedi. "Reading with sounds: sensory substitution selectively activates the visual word form area in the blind". In: *Neuron* 76.3 (2012), pp. 640–652.

[13] Yoeri M Luijf, Julia K Mader, Werner Doll, Thomas Pieber, Anne Farret, Jerome Place, Eric Renard, Daniela Bruttomesso, Alessio Filippi, Angelo Avogaro, *et al.* "Accuracy and reliability of continuous glucose monitoring systems: a head-to-head comparison". In: *Diabetes technology & therapeutics* 15.8 (2013), pp. 721–726.

[14] Paul Bach-y-Rita, Mitchell E Tyler, and Kurt A Kaczmarek. "Seeing with the brain". In: *International journal of human-computer interaction* 15.2 (2003), pp. 285–295.

[15] Ladan Shams and Robyn Kim. "Crossmodal influences on visual perception". In: *Physics of Life Reviews* 7.3 (2010), pp. 269–284.

[16] Gemma A Calvert, Michael J Brammer, and Susan D Iversen. "Crossmodal identification". In: *Trends in Cognitive Sciences* 2.7 (1998), pp. 247–253.

[17] Gemma A Calvert, Michael J Brammer, Edward T Bullmore, Ruth Campbell, Susan D Iversen, and Anthony S David. “Response amplification in sensory-specific cortices during crossmodal binding”. In: *Neuroreport* 10.12 (1999), pp. 2619–2623.

[18] Asif A Ghazanfar and Charles E Schroeder. “Is neocortex essentially multisensory?” In: *Trends in cognitive sciences* 10.6 (2006), pp. 278–285.

[19] Ian P Howard and William B Templeton. *Human spatial orientation*. Wiley, 1966.

[20] David H Warren, Robert B Welch, and Thomas J McCarthy. “The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses”. In: *Perception Psychophysics* 30.6 (1981), pp. 557–564.

[21] Irvin Rock and James Victor. “Vision and touch: An experimentally created conflict between the two senses”. In: *Science* 143.3606 (1964), pp. 594–596.

[22] Charles Spence and Jon Driver, eds. *Crossmodal space and crossmodal attention*. Oxford University Press, 2004.

[23] Harry McGurk and John MacDonald. “Hearing lips and seeing voices”. In: *Nature* 264.5588 (1976), pp. 746–748.

[24] Robert Sekuler, Allison B Sekuler, and Rene Lau. “Sound alters visual motion perception”. In: *Nature* 385.6614 (1997), p. 308.

[25] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. “What you see is what you hear”. In: *Nature* 408.6814 (2000), p. 788.

[26] Charles Spence. “Crossmodal correspondences: A tutorial review”. In: *Attention, Perception, Psychophysics* 73.4 (2011), pp. 971–995.

[27] Marc O Ernst. “Learning to integrate arbitrary signals from vision and touch”. In: *Journal of Vision* 7.5 (2007), p. 7.

[28] Marc O Ernst and Heinrich H B"ulthoff. “Merging the senses into a robust percept”. In: *Trends in Cognitive Sciences* 8.4 (2004), pp. 162–169.

[29] Neil W Roach, James Heron, and Paul V McGraw. “Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration”. In: *Proceedings of the Royal Society B: Biological Sciences* 273.1598 (2006), pp. 2159–2168.

[30] Konrad P K"ording, Ulrik Beierholm, Wei Ji Ma, Joshua B Tenenbaum, Steven Quartz, and Ladan Shams. “Causal inference in multisensory perception”. In: *PLoS One* 2.9 (2007), e943.

[31] Jack A Jones and Michelle Jarick. “Crossmodal congruence effects in speeded classification tasks”. In: *Experimental Brain Research* 174.3 (2006), pp. 387–396.

[32] David I Shore, Melody E Barnes, and Charles Spence. “The temporal evolution of the crossmodal congruency effect”. In: *Neuroscience Letters* 392.1 (2006), pp. 96–100.

[33] Gemma A Calvert, Charles Spence, and Barry E Stein. “The handbook of multisensory processes”. In: MIT Press, 2004.

[34] Paul Bertelson, Jean Vroomen, Geert Wiegeraad, and Beatrice de Gelder. “Ventriloquism and audiovisual interaction”. In: *Proceedings of the 1994 International Conference on Spoken Language Processing* 2 (1994), pp. 559–562.

[35] Hamish Innes-Brown and David Crewther. “The impact of spatial incongruence on an auditory–visual illusion”. In: *PloS one* 4.7 (2009), e6450.

[36] Yi-Chuan Chen and Charles Spence. “Semantically congruent visual information enhances processing of auditory objects”. In: *Cognition* 114.3 (2010), pp. 389–404.

[37] Paul J Laurienti, Mark T Wallace, Joseph A Maldjian, Christine M Susi, Barry E Stein, and Jonathan H Burdette. “Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices”. In: *Human Brain Mapping* 19.4 (2003), pp. 213–223.

[38] Sarah Jessen and Sonja A Kotz. “On the role of crossmodal prediction in audiovisual emotion perception”. In: *Frontiers in Human Neuroscience* 7 (2013), p. 369.

[39] Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-Le’on. “Music–color associations are mediated by emotion”. In: *Proceedings of the National Academy of Sciences* 110.22 (2013), pp. 8836–8841.

[40] David Alais and David Burr. “Bimodal integration of vision and touch”. In: *Current Biology* 14 (2004), pp. 257–262.

[41] Alvaro Pascual-Leone, Amir Amedi, Felipe Fregni, and Lotfi B Merabet. “Visual deprivation and cross-modal plasticity”. In: *Annual Review of Neuroscience* 28 (2005), pp. 377–401.

[42] Lotfi B Merabet, Patrice Voss, Maryse Lassonde, and Franco Lepore. “Functional specialization for auditory-spatial processing in the occipital cortex of congenitally blind humans”. In: *Proceedings of the National Academy of Sciences* 104.16 (2007), pp. 4429–4434.

[43] Helen J Neville and Daphne Bavelier. “Cerebral organization for language in deaf and hearing subjects: biological constraints and effects of experience”. In: *Proceedings of the National Academy of Sciences* 95.3 (1998), pp. 922–929.

[44] Olivier Collignon, Patrice Voss, Maryse Lassonde, and Franco Lepore. “Functional specialization for auditory-spatial processing in the occipital cortex of congenitally blind humans”. In: *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 4429–4434.

[45] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. “Statistical learning by 8-month-old infants”. In: *Science* 274.5294 (1996), pp. 1926–1928.

[46] Cesare V Parise and Charles Spence. “"When birds of a feather flock together": Synesthetic correspondences modulate audiovisual integration in non-synesthetes”. In: *PloS one* 4.5 (2009), e5664.

[47] Leonid Sabaneev and SW Pring. “The relation between sound and colour”. In: *Music & Letters* 10.3 (1929), pp. 266–277.

[48] Theodore F Karwoski and Henry S Odbert. “Color-music.” In: *Psychological Monographs* 50.2 (1938), p. i.

[49] Henry S Odbert, Theodore F Karwoski, and AB Eckerson. “Studies in synesthetic thinking: I. Musical and verbal associations of color and mood”. In: *The journal of general psychology* 26.1 (1942), pp. 153–173.

[50] Chris J Boyatzis and Reenu Varghese. “Children’s emotional associations with colors”. In: *The Journal of genetic psychology* 155.1 (1994), pp. 77–85.

[51] Michael Hemphill. “A note on adults’ color–emotion associations”. In: *The Journal of genetic psychology* 157.3 (1996), pp. 275–280.

[52] Meyer Leonard. “Emotion and meaning in music”. In: *Chicago: University of Chicago* (1956).

[53] Paul J Laurienti, Robert A Kraft, Joseph A Maldjian, Jonathan H Burdette, and Mark T Wallace. “Semantic congruence is a critical factor in multisensory behavioral performance”. In: *Experimental brain research* 158 (2004), pp. 405–414.

[54] Lawrence E Marks. “Cross-modal interactions in speeded classification.” In: (2004).

[55] Roberto Bresin. “What is the color of that music performance?” In: *ICMC*. 2005.

[56] J Michael Barbiere, Ana Vidal, and Debra A Zellner. “The color of music: Correspondence through emotion”. In: *Empirical studies of the arts* 25.2 (2007), pp. 193–208.

[57] Costanza Cenerini, Luca Vollero, Giorgio Pennazza, Marco Santonico, Nicola Di Stefano, and Flavio Keller. “INVESTIGATING COLOUR-SOUND MAPPING IN CHILDREN AND ADULTS: A PILOT STUDY”. In: *Proceedings of ICAD Conference* (2023).

[58] Thomas Hermann, Andy Hunt, and John G Neuhoff. *The Sonification Handbook*. Logos Verlag Berlin, 2011.

[59] Richard Parncutt. *Harmony: A Psychoacoustical Approach*. Springer-Verlag, 1989.

[60] Catherine J Mondloch and Daphne Maurer. “Do small white balls squeak? Pitch-object correspondences in young children”. In: *Cognitive, Affective, Behavioral Neuroscience* 4.2 (2004), pp. 133–136.

[61] Peter Walker, J Gavin Bremner, Uschi Mason, Jo Spring, Karen Mattock, Alan Slater, and Scott P Johnson. “Preverbal infants’ sensitivity to synaesthetic cross-modality correspondences”. In: *Psychological Science* 21.1 (2010), pp. 21–25.

[62] Charles Spence and Nicola Di Stefano. “Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and sound”. In: *i-Perception* 13.3 (2022), p. 20416695221092802.

[63] A. Lavignac. *Music and musicians*. Henry Holt and Company, 1899.

[64] Zachary Wallmark. “Semantic crosstalk in timbre perception”. In: *Music & Science* 2 (2019), p. 2059204319846617.

[65] J. W. von Goethe. *Theory of colours*. Trans. by Charles Locke Eastlake. Original work published 1810. John Murray, 1840.

[66] H. von Helmholtz. *Handbuch der Physiologischen Optik [Handbook of Physiological Optics]*. Voss, 1867.

[67] Andrey Anikin and Niklas Johansson. "Implicit associations between individual properties of color and sound". In: *Attention, Perception, & Psychophysics* 81 (2019), pp. 764–777.

[68] Lawrence E Marks. "On cross-modal similarity: Auditory–visual interactions in speeded discrimination." In: *Journal of experimental psychology: Human Perception and Performance* 13.3 (1987), p. 384.

[69] Lawrence E Marks. "On cross-modal similarity: the perceptual structure of pitch, loudness, and brightness." In: *Journal of Experimental Psychology: Human Perception and Performance* 15.3 (1989), p. 586.

[70] Bulat M Galeyev and Irina L Vanechkina. "Was Scriabin a synesthete?" In: *Leonardo* 34.4 (2001), pp. 357–361.

[71] Costanza Cenerini, Luca Vollero, Giorgio Pennazza, Marco Santonico, and Flavio Keller. "Audio Visual Association Test in Non Synesthetic Subjects: Technological Tailoring of the Methods". In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer. 2022, pp. 432–437.

[72] Paul Ekman. "Are there basic emotions?" In: (1992).

[73] Julian Cespedes-Guevara and Tuomas Eerola. "Music communicates affects, not basic emotions–A constructionist account of attribution of emotional meanings to music". In: *Frontiers in psychology* 9 (2018), p. 215.

[74] Anna Aljanaki, Frans Wiering, Remco Veltkamp, *et al.* "Computational modeling of induced emotion using GEMS". In: *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014)*. 2014, pp. 373–378.

[75] Joshua Klayman. "Varieties of confirmation bias". In: *Psychology of learning and motivation* 32 (1995), pp. 385–418.

[76] Peter MC Harrison, Jason Jiří Musil, and Daniel Müllensiefen. "Modelling melodic discrimination tests: Descriptive and explanatory approaches". In: *Journal of New Music Research* 45.3 (2016), pp. 265–280.

[77] Pauline Larrouy-Maestri, Peter Harrison, and Daniel Müllensiefen. "The mistuning perception test: A new measurement instrument". In: *Behavior Research Methods* 51.2 (2019), pp. 663–675.

[78] Peter Harrison and Daniel Müllensiefen. “Development and validation of the computerised adaptive beat alignment test (CA-BAT)”. In: *Scientific Reports* 8.1 (2018), pp. 1–19.

[79] Morris K Holland and Michael Wertheimer. “Some physiognomic aspects of naming, or, maluma and takete revisited”. In: *Perceptual and Motor Skills* 19.1 (1964), pp. 111–117.

[80] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. “A matlab toolbox for music information retrieval”. In: *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation eV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. Springer Berlin Heidelberg. 2008.

[81] Fiorenzo Conti. “Fisiologia delle emozioni”. In: *Fisiologia Medica*. 3rd ed. Milano: Edi Ermes, 2019. Chap. 35, pp. 781–802.

[82] Paul Ekman. “An argument for basic emotions”. In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[83] Paul Ekman and Wallace V Friesen. “Cross-cultural studies of facial expression”. In: *Darwin and facial expression: A century of research in review* (1973), pp. 169–222.

[84] Paul Ekman and Wallace V Friesen. “Constants across cultures in the face and emotion”. In: *Journal of personality and social psychology* 17.2 (1971), p. 124.

[85] Paul Ekman, Wallace V Friesen, Maureen O’Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, *et al.* “Universals and cultural differences in the judgments of facial expressions of emotion”. In: *Emotion and social behavior*. Ed. by Paul Ekman. Sage, 1987, pp. 712–717.

[86] Paul Ekman. “Strong evidence for universals in facial expressions: A reply to Russell’s mistaken critique”. In: *Psychological Bulletin* 115.2 (1994), pp. 268–287.

[87] Carroll E Izard. “The face of emotion”. In: (1971).

[88] Rafael A Calvo and Sidney D’Mello. “Affect detection: An interdisciplinary review of models, methods, and their applications”. In: *IEEE Transactions on affective computing* 1.1 (2010), pp. 18–37.

[89] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. "Multimodal approaches for emotion recognition: a survey". In: *Proceedings of SPIE* 5670 (2005), pp. 56–67.

[90] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[91] Margaret M Bradley, Mark K Greenwald, Margaret C Petry, and Peter J Lang. "Remembering pictures: pleasure and arousal in memory". In: *Journal of experimental psychology: Learning, Memory, and Cognition* 18.2 (1992), p. 379.

[92] Peter J Lang. "The emotion probe: studies of motivation and attention". In: *American psychologist* 50.5 (1995), p. 372.

[93] Albert Mehrabian. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". In: *Current Psychology* 14.4 (1996), pp. 261–292.

[94] Jonathan Posner, James A Russell, and Bradley S Peterson. "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology". In: *Development and psychopathology* 17.3 (2005), pp. 715–734.

[95] Patrik N Juslin and John A Sloboda. "Music and emotion: Theory and research". In: (2001).

[96] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. "Emotions evoked by the sound of music: characterization, classification, and measurement". In: *Emotion* 8.4 (2008), p. 494.

[97] Tuomas Eerola and Jonna K Vuoskoski. "A comparison of the discrete and dimensional models of emotion in music". In: *Psychology of Music* 39.1 (2010), pp. 18–49.

[98] Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[99] Susan Folkman, Richard S Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J Gruen. "Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes". In: *Journal of personality and social psychology* 50.5 (1986), p. 992.

[100] Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

[101] Andrew Ortony, Gerald L Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press, 1988.

[102] Christoph Bartneck. "Integrating the OCC Model of Emotions in Embodied Characters". In: *Proceedings of the Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges*. Melbourne, 2002.

[103] Walter B. Cannon. "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory". In: *The American Journal of Psychology* 39.1/4 (1927), pp. 106–124.

[104] William James. "What is an emotion?" In: *Mind* 9.34 (1884), pp. 188–205.

[105] Antonio R Damasio. "The somatic marker hypothesis and the possible functions of the prefrontal cortex". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351.1346 (1996), pp. 1413–1420.

[106] Harald Kindermann and Franz Auinger. "Emotions and Feelings: Some Aspects for the HCI-Community–A Work in Progress Paper". In: *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 5*. Springer. 2018, pp. 338–350.

[107] Alessia Celeghin, Matteo Diano, Arianna Bagnis, Marco Viola, and Marco Tamietto. "Basic emotions in human neuroscience: neuroimaging and beyond". In: *Frontiers in psychology* 8 (2017), p. 1432.

[108] Luiz Pessoa. "On the relationship between emotion and cognition". In: *Nature reviews neuroscience* 9.2 (2008), pp. 148–158.

[109] Ala Yankouskaya, Moritz Stolte, Zargol Moradi, Pia Rotshtein, and Jie Sui. "Neural connectivity underlying reward and emotion-related processing: Evidence from a large-scale network analysis". In: *Frontiers in systems neuroscience* 16 (2022), p. 833625.

[110] Elenor Morgenroth, Ulrike Kr"amer, Martin G"ottlich, Julia Br"andle, Anna Michalska, Sarah Alizadeh, Tessa Rosenkranz, Nils Onken, and Martin Lotze. "Probing neurodynamics of experienced emotions—a Hitchhiker's guide to film fMRI". In: *Social Cognitive and Affective Neuroscience* 18.1 (2023), nsad063.

[111] Chunliang Feng, Cui Xiao, and Xiujun Zhou. “Separate neural networks of implicit emotional processing between pictures and words: a coordinate-based meta-analysis of brain imaging studies”. In: *Neuroscience Biobehavioral Reviews* 131 (2021), pp. 331–344.

[112] Suzanne N Haber, Anastasia Yendiki, and Saad Jbabdi. “Prefrontal connectomics: from anatomy to human imaging”. In: *Neuropsychopharmacology* 47.1 (2022), pp. 20–40.

[113] Miranda Wood, Muhammed Adil, Michael Gravier, Sonam Dolma, and P Jeffrey Conn. “Infralimbic prefrontal cortex structural and functional connectivity with the limbic forebrain: a combined viral genetic and optogenetic analysis”. In: *Brain Structure and Function* 224 (2019), pp. 73–97.

[114] Yixiao Fu, Zhi Li, Zhenglong Xu, Mingjun Duan, Zhengyan Yu, Yuxiao Zhao, Weiwei Wu, Panpan Hu, Xiaocui Zhang, Qiong Yuan, *et al.* “Functional and structural connectivity between the left dorsolateral prefrontal cortex and insula could predict the antidepressant effects of repetitive transcranial magnetic stimulation”. In: *Frontiers in neuroscience* 15 (2021), p. 645936.

[115] Karine Sergerie, Caroline Chochol, and Jorge L Armony. “The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies”. In: *Neuroscience Biobehavioral Reviews* 32.4 (2008), pp. 811–830.

[116] Stefan Koelsch, Thomas Fritz, D Yves v Cramon, Karsten M"uller, and Angela D Friederici. “Investigating emotion with music: an fMRI study”. In: *Human brain mapping* 27.3 (2006), pp. 239–250.

[117] Zhenhong He, Ruida Zhou, Xinghua Li, Zheming Huang, Yunhong Xiang, Rui Kong, Yue Zhang, Jialin Wang, Jilei Zhang, Jiang Zhang, *et al.* “The VLPFC-engaged voluntary emotion regulation: Combined TMS-fMRI evidence for the neural circuit of cognitive reappraisal”. In: *Journal of Neuroscience* 43.34 (2023), pp. 6046–6060.

[118] Micha Keller, Jana Zweerings, Martin Klasen, Armin Karpinski, Sarah Alizadeh, Joseph Kambeitz, Bj"orn H Schott, Joern Kaufmann, Sigrid Elsenbruch, Peter Zwanzger, *et al.* “Transdiagnostic alterations in neural emotion regulation circuits–neural substrates of cognitive reappraisal in patients with depression and post-traumatic stress disorder”. In: *BMC psychiatry* 22.1 (2022), p. 173.

[119] Sarah Opialla, Jacqueline Lutz, Sigrid Scherpiet, Anna Hittmeyer, Lutz J"ancke, Michael Rufer, Martin Grosse Holtforth, Uwe Herwig, and Annette B Br"uhl. “Neural circuits of emotion regulation: A comparison of mindfulness-based and cognitive reappraisal strategies.” In: *European archives of psychiatry and clinical neuroscience* 265.1 (2015), pp. 45–55.

[120] Sean T Ma, James L Abelson, Go Okada, Stephan F Taylor, and Israel Liberzon. “Neural circuitry of emotion regulation: Effects of appraisal, attention, and cortisol administration”. In: *Cognitive, Affective, Behavioral Neuroscience* 17 (2017), pp. 437–451.

[121] Luiz Pessoa and Leslie G Ungerleider. “Neuroimaging studies of attention and the processing of emotion-laden stimuli”. In: *Progress in brain research* 144 (2004), pp. 171–182.

[122] Matthias Schurz, Marlene Hackl, Markus Aichhorn, Teresa Simbrunner, Martin Kronbichler, and Josef Perner. “Brain Activation for Social Cognition and Emotion Processing Tasks in Borderline Personality Disorder: A Meta-Analysis of Neuroimaging Studies”. In: *Brain Sciences* 14.4 (2024), p. 395.

[123] Jared Rieck, Emma T McKinnon, Joseph W Kable, Michael Roy, and Mandeep K Singh. “Neural signatures of emotion regulation”. In: *Scientific reports* 14.1 (2024), p. 1775.

[124] Yoshiya Moriguchi and Gen Komaki. “Neuroimaging studies of alexithymia: physical, affective, and social perspectives”. In: *BioPsychoSocial medicine* 7 (2013), pp. 1–12.

[125] Kathryn Berluti, Montana L Ploe, and Abigail A Marsh. “Emotion processing in youths with conduct problems: an fMRI meta-analysis”. In: *Translational psychiatry* 13.1 (2023), p. 105.

[126] Rosalind W Picard. “Affective Computing”. In: *M.I.T Media Laboratory Perceptual Computing Section Technical Report* 321 (1995), pp. 1–16.

[127] Rosalind W Picard. *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[128] Manh-Tung Ho *et al.* “Affective computing scholarship and the rise of China: a view from 25 years of bibliometric data”. In: *Humanities and Social Sciences Communications* 8.1 (2021), pp. 1–14.

[129] Guanxiong Pei *et al.* "Affective Computing: Recent Advances, Challenges, and Future Trends". In: *Intelligent Computing* 3 (2024), p. 0076.

[130] Haiwei Ma and Svetlana Yarosh. "A review of affective computing research based on function-component-representation framework". In: *IEEE Transactions on Affective Computing* 14.2 (2021), pp. 1655–1674.

[131] Yiqun Zhang *et al.* "Affective computing in the era of large language models: A survey from the NLP perspective". In: *arXiv preprint arXiv:2408.04638* (2024).

[132] Björn Schuller *et al.* "Affective Computing Has Changed: The Foundation Model Disruption". In: *arXiv preprint arXiv:2409.08907* (2024).

[133] Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors". In: *Scientific Reports* 8.1 (2018), p. 13657.

[134] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, *et al.* "International affective picture system (IAPS): Technical manual and affective ratings". In: *NIMH Center for the Study of Emotion and Attention* 1.39-58 (1997), p. 3.

[135] Ryan A Stevenson and Thomas W James. "Affective auditory stimuli: Characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories". In: *Behavior research methods* 40.1 (2008), pp. 315–321.

[136] Margaret M Bradley and Peter J Lang. "Measuring emotion: the self-assessment manikin and the semantic differential". In: *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), pp. 49–59.

[137] Laurence Devillers and Roddy Cowie. "Ethical considerations on affective computing: an overview". In: *Proceedings of the IEEE* (2023).

[138] Bradley M Appelhans and Linda J Luecken. "Heart rate variability as an index of regulated emotional responding". In: *Review of general psychology* 10.3 (2006), pp. 229–240.

[139] Hui Wen Loh, Shuting Xu, Oliver Faust, Chui Ping Ooi, Prabal Datta Barua, Subrata Chakraborty, Ru-San Tan, Filippo Molinari, and U Rajendra Acharya. "Application of photoplethysmography signals for healthcare systems: An in-depth review". In: *Computer Methods and Programs in Biomedicine* 216 (2022), p. 106677.

[140] Sylvia D Kreibig. "Autonomic nervous system activity in emotion: A review". In: *Biological psychology* 84.3 (2010), pp. 394–421.

[141] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

[142] Anton Van Boxtel. "Facial EMG as a tool for inferring affective states". In: *Proceedings of measuring behavior* 2010 (2010), pp. 104–108.

[143] Anton Van Boxtel. "Facial EMG as a tool for inferring affective states". In: *Proceedings of measuring behavior*. Vol. 2010. 2010.

[144] Jeff T Larsen, Catherine J Norris, and John T Cacioppo. "Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii". In: *Psychophysiology* 40.5 (2003), pp. 776–785.

[145] Gary D James, Lori S Yee, Gregory A Harshfield, Steven G Blank, and Thomas G Pickering. "Cardiovascular changes during induced emotion: An application of Lang's theory of emotional imagery". In: *Journal of psychosomatic research* 75.2 (2013), pp. 132–138.

[146] Fons A Boiten, Nico H Frijda, and Cornelis JE Wientjes. "Emotions and respiratory patterns: review and critical analysis". In: *International Journal of Psychophysiology* 28.2 (1998), pp. 121–140.

[147] Stéphanie Khalfa, Simone Dalla Bella, Mathieu Roy, Isabelle Peretz, and Sonia J Lupien. "Effects of relaxing music on salivary cortisol level after psychological stress". In: *Annals of the New York Academy of Sciences* 999.1 (2002), pp. 374–376.

[148] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. "Looking at pictures: Affective, facial, visceral, and behavioral reactions". In: *Psychophysiology* 30.3 (1993), pp. 261–273.

[149] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. "A multimodal database for affect recognition and implicit tagging". In: *IEEE transactions on affective computing* 3.1 (2012), pp. 42–55.

[150] Brais Martinez and Michel F Valstar. “Advances, challenges, and opportunities in automatic facial expression recognition”. In: *Advances in face detection and facial image analysis* (2016), pp. 63–100.

[151] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. “Meta-analysis of the first facial expression recognition challenge”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 966–979.

[152] I Michael Revina and WR Sam Emmanuel. “A survey on human face expression recognition techniques”. In: *Journal of King Saud University-Computer and Information Sciences* 33.6 (2021), pp. 619–628.

[153] Paul Ekman *et al.* “Basic emotions”. In: *Handbook of cognition and emotion* 98.45-60 (1999), p. 16.

[154] J Anil and L Padma Suresh. “Literature survey on face and face expression recognition”. In: *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE. 2016, pp. 1–6.

[155] Paul Ekman and Wallace V Friesen. “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior* (1978).

[156] JH Cheong, T Xie, S Byrne, and LJ Chang. *Py-Feat: Python facial expression analysis toolbox (arXiv: 2104.03509)*. *arXiv*. 2021.

[157] Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. “Emotion detection: a technology review”. In: *Proceedings of the XVIII international conference on human computer interaction*. 2017, pp. 1–8.

[158] Ankita Verma, Dhutima Malla, Amrit Kaur Choudhary, and Vasudha Arora. “A detailed study of azure platform & its cognitive services”. In: *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE. 2019, pp. 129–134.

[159] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. “AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit”. In: *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 2016, pp. 3723–3726.

[160] Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C Samson. "Facial expression analysis with AFFDEX and FACET: A validation study". In: *Behavior research methods* 50 (2018), pp. 1446–1460.

[161] Martin Magdin, L'ubomír Benko, and Štefan Koprda. "A case study of facial emotion classification using affdex". In: *Sensors* 19.9 (2019), p. 2140.

[162] Justadudewhohacks. *Justadudewhohacks/face-api.js: JavaScript API for face detection and face recognition in the browser and nodejs with tensorflow.js*. URL: https://github.com/justadudewhohacks/face-api.js/.

[163] Byoung Chul Ko. "A brief review of facial emotion recognition based on visual information". In: *sensors* 18.2 (2018), p. 401.

[164] Yunxin Huang, Fei Chen, Shaohe Lv, and Xiaodong Wang. "Facial expression recognition: A survey". In: *Symmetry* 11.10 (2019), p. 1189.

[165] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild". In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[166] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression". In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE. 2010, pp. 94–101.

[167] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. "Comprehensive database for facial expression analysis". In: *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*. IEEE. 2000, pp. 46–53.

[168] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. "Deep learning approaches for facial emotion recognition: A case study on FER-2013". In: *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications* (2018), pp. 1–16.

[169] Saranya Rajan, Poongodi Chenniappan, Somasundaram Devaraj, and Nirmala Madian. "Facial expression recognition techniques: a comprehensive survey". In: *IET Image Processing* 13.7 (2019), pp. 1031–1040.

[170] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. “Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.” In: *2003 Conference on computer vision and pattern recognition workshop*. Vol. 5. IEEE. 2003, pp. 53–53.

[171] Andrea Kleinsmith and Nadia Bianchi-Berthouze. “Affective body expression perception and recognition: A survey”. In: *IEEE Transactions on Affective Computing* 4.1 (2013), pp. 15–33.

[172] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. “Body movements for affective expression: A survey of automatic recognition and generation”. In: *IEEE Transactions on Affective Computing* 4.4 (2013), pp. 341–359.

[173] Sidney K D’mello and Jacqueline Kory. “A review and meta-analysis of multimodal affect detection systems”. In: *ACM Computing Surveys (CSUR)* 47.3 (2015), pp. 1–36.

[174] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. “A review of affective computing: From unimodal analysis to multimodal fusion”. In: *Information Fusion* 37 (2017), pp. 98–125.

[175] Mehmet Berke Akçay and Kaya Oğuz. “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”. In: *Speech Communication* 116 (2020), pp. 56–76.

[176] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. “Survey on speech emotion recognition: Features, classification schemes, and databases”. In: *Pattern Recognition* 44.3 (2011), pp. 572–587.

[177] Ala Saleh Alluhaidan, Oumaima Saidani, Rashid Jahangir, Muhammad Asif Nauman, and Omnia Saidani Neffati. “Speech emotion recognition through hybrid features and convolutional neural network”. In: *Applied Sciences* 13.8 (2023), p. 4750.

[178] Björn W Schuller. “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends”. In: *Communications of the ACM* 61.5 (2018), pp. 90–99.

[179] Ludovica La Monica, Costanza Cenerini, Luca Vollero, Giorgio Pennazza, Marco Santonico, and Flavio Keller. "Development of a Universal Validation Protocol and an Open-Source Database for Multi-Contextual Facial Expression Recognition". In: *Sensors* 23.20 (2023), p. 8376.

[180] Costanza Cenerini. "Towards affective sensory substitution". In: *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. ©2024 IEEE. Glasgow, UK: IEEE, Sept. 2024, pp. 92–96. ISBN: 979-8-3315-1645-1. DOI: 10.1109/ACIIW63320.2024.00019.

[181] Benjamin Kreifelts, Thomas Ethofer, Wolfgang Grodd, Michael Erb, and Dirk Wildgruber. "Audiovisual integration of emotional signals in voice and face: An event-related fMRI study". In: *NeuroImage* 37.4 (2007), pp. 1445–1456.

[182] Philip A Kragel, Marianne C Reddan, Kevin S LaBar, and Tor D Wager. "Emotion schemas are embedded in the human visual system". In: *Science advances* 5.7 (2019), eaaw4358.

[183] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5562–5570.

[184] Hayette Hadjar, Thoralf Reis, Marco X Bornschlegl, Felix C Engel, Paul Mc Kevitt, and Matthias L Hemmje. "Recognition and visualization of facial expression and emotion in healthcare". In: *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications: AVI 2020 Workshops, AVI-BDA and ITAVIS, Ischia, Italy, June 9, 2020 and September 29, 2020, Revised Selected Papers*. Springer. 2021, pp. 109–124.

[185] Eylül Ertay, Hao Huang, Zhanna Sarsenbayeva, and Tilman Dingler. "Challenges of emotion detection using facial expressions and emotion visualisation in remote communication". In: *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 2021, pp. 230–236.

[186] Dongchun Ren, Peng Wang, Hong Qiao, and Suiwu Zheng. “A biologically inspired model of emotion eliciting from visual stimuli”. In: *Neurocomputing* 121 (2013), pp. 328–336.

[187] Daniel Grühn and Neika Sharifian. “Lists of emotional stimuli”. In: *Emotion measurement*. Elsevier, 2016, pp. 145–164.

[188] Bryn Farnsworth, Divya Seernani, Pernille Bülow, and Kate Krosschell. *The International Affective Picture System [explained and alternatives]*. Nov. 2022. URL: https://imotions.com/blog/learning/research-fundamentals/iaps-international-affective-picture-system/.

[189] Michela Balsamo, Leonardo Carlucci, Caterina Padulo, Bernardo Perfetti, and Beth Fairfield. “A bottom-up validation of the IAPS, GAPED, and NAPS affective picture databases: Differential effects on behavioral performance”. In: *Frontiers in Psychology* 11 (2020), p. 2187.

[190] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. “Introducing the open affective standardized image set (OASIS)”. In: *Behavior research methods* 49 (2017), pp. 457–470.

[191] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. “The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database”. In: *Behavior research methods* 46 (2014), pp. 596–610.

[192] Monika Riegel, Łukasz Żurawski, Małgorzata Wierzba, Abnoss Moslehi, Łukasz Klocek, Marko Horvat, Anna Grabowska, Jarosław Michałowski, Katarzyna Jednoróg, and Artur Marchewka. “Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE)”. In: *Behavior research methods* 48 (2016), pp. 600–612.

[193] Pilar Ferré, Marc Guasch, Cornelia Moldovan, and Rosa Sánchez-Casas. “Affective norms for 380 Spanish words belonging to three different semantic categories”. In: *Behavior Research Methods* 44 (2012), pp. 395–403.

[194] Johanna Kissler, Cornelia Herbert, Peter Peyk, and Markus Junghofer. “Buzzwords: early cortical responses to emotional words during reading”. In: *Psychological science* 18.6 (2007), pp. 475–480.

[195] James A Russell and Albert Mehrabian. “Evidence for a three-factor theory of emotions”. In: *Journal of research in Personality* 11.3 (1977), pp. 273–294.

[196] James A Russell and Lisa Feldman Barrett. “Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.” In: *Journal of personality and social psychology* 76.5 (1999), p. 805.

[197] Divya Garg, Gyanendra Kumar Verma, and Awadhesh Kumar Singh. “Modelling and statistical analysis of emotions in 3D space”. In: *Engineering Research Express* 4.3 (2022), p. 035062.

[198] Marko Horvat, Alan Jović, and Kristijan Burnik. “Investigation of relationships between discrete and dimensional emotion models in affective picture databases using unsupervised machine learning”. In: *Applied Sciences* 12.15 (2022), p. 7864.

[199] Sieun An, Li-Jun Ji, Michael Marks, and Zhiyong Zhang. “Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures”. In: *Frontiers in psychology* 8 (2017), p. 610.

[200] EA Yumatov. “Duality of the Nature of Emotions and Stress: Neurochemical Aspects”. In: *Neurochemical Journal* 16.4 (2022), pp. 429–442.

[201] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. “A review on automatic facial expression recognition systems assisted by multimodal sensor data”. In: *Sensors* 19.8 (2019), p. 1863.

[202] Shan Li and Weihong Deng. “Deep facial expression recognition: A survey”. In: *IEEE transactions on affective computing* 13.3 (2020), pp. 1195–1215.

[203] Vollero Luca, Cenerini Costanza, and La Monica Ludovica. *FeelPix [Landmark database]*. GitHub repository: *https://github.com/ludovicalamonica/Feel* 2023.

[204] Matthias Schmidmaier. “Sensory substitution systems”. In: *Media Informatics Advanced Seminar on Multimodal Human-Computer Interaction*. 2011.

[205] David M Eagleman and Michael V Perrotta. “The future of sensory substitution, addition, and expansion via haptic devices”. In: *Frontiers in Human Neuroscience* 16 (2023), p. 1055546.

[206] Amy C Nau, Matthew C Murphy, and Kevin C Chan. “Use of sensory substitution devices as a model system for investigating cross-modal neuroplasticity in humans”. In: *Neural regeneration research* 10.11 (2015), pp. 1717–1719.

[207] Paul Bach-y-Rita, Carter C Collins, Frank A Saunders, Benjamin White, and Lawrence Scadden. “Vision substitution by tactile image projection”. In: *Nature* 221.5184 (1969), pp. 963–964.

[208] Susan Hurley and Alva Noe. “Neural plasticity and consciousness: reply to block”. In: *Trends in cognitive sciences* 7.8 (2003), p. 342.

[209] Paul Bach-y-Rita, Yuri Danilov, Mitchell Tyler, and Robert Grimm. “Late human brain plasticity: vestibular substitution with a tongue BrainPort human-machine interface”. In: *Intellectica* 40.1 (2005), pp. 115–122.

[210] Amy C Nau, Christine Pintar, Aimee Arnoldussen, and Christopher Fisher. “Acquisition of visual perception in blind adults using the BrainPort artificial vision device”. In: *The American Journal of Occupational Therapy* 69.1 (2015), 6901290010p1–6901290010p8.

[211] Peter BL Meijer. “An experimental system for auditory image representations”. In: *IEEE transactions on biomedical engineering* 39.2 (1992), pp. 112–121.

[212] Sami Abboud, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi. “EyeMusic: Introducing a “visual” colorful experience for the blind using auditory sensory substitution”. In: *Restorative neurology and neuroscience* 32.2 (2014), pp. 247–257.

[213] Izzy Kohler, Michael V Perrotta, Tiago Ferreira, and David M Eagleman. “Cross-modal sensory boosting to improve high-frequency hearing loss: device development and validation”. In: *JMIRx Med* 5 (2024), e49969.

[214] Rebekka Hoffmann, Simone Spagnol, Árni Kristjánsson, and Runar Unnthorsson. “Evaluation of an audio-haptic sensory substitution device for enhancing spatial awareness for the visually impaired”. In: *Optometry and Vision Science* 95.9 (2018), pp. 757–765.

[215] Oran Goral *et al.* “Enhancing interoceptive sensibility through exteroceptive–interoceptive sensory substitution”. In: *Scientific Reports* 14.1 (2024), p. 14855.

[216] Christopher Festin *et al.* "Creation of a biological sensorimotor interface for bionic reconstruction". In: *Nature Communications* 15.1 (2024), p. 5337.

[217] Patrik N Juslin and John Sloboda. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2011.

[218] William Forde Thompson. *Music, thought, and feeling: Understanding the psychology of music*. Oxford university press, 2015.

[219] Stefan Koelsch. "Brain correlates of music-evoked emotions". In: *Nature Reviews Neuroscience* 15.3 (2014), pp. 170–180.

[220] Tuomas Eerola and Jonna K Vuoskoski. "A review of music and emotion studies: Approaches, emotion models, and stimuli". In: *Music Perception* 30.3 (2013), pp. 307–340.

[221] Laura-Lee Balkwill and William Forde Thompson. "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues". In: *Music Perception* 17.1 (1999), pp. 43–64.

[222] Patricia Valdez and Albert Mehrabian. "Effects of color on emotions." In: *Journal of experimental psychology: General* 123.4 (1994), p. 394.

[223] Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-Le'on. "Music–color associations are mediated by emotion". In: *Proceedings of the National Academy of Sciences* 110.22 (2013), pp. 8836–8841.

[224] Xiaowei Lu, Pabitra Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. "Shape theory based emotional expression recognition in art paintings". In: *IEEE Transactions on Multimedia* 14.5 (2012), pp. 1431–1441.

[225] Jana Machajdik and Allan Hanbury. "Affective image classification using features inspired by psychology and art theory". In: *Proceedings of the 18th ACM International Conference on Multimedia* (2010), pp. 83–92.

[226] Dandan Zhang, Yuqin Liu, Chenggang Zhou, Yuping Chen, and Yunxiang Luo. "Temporal dynamics of audiovisual affective processing". In: *Biological Psychology* 147 (2019), pp. 107–119.

[227] Xiaohui Wang, Lin Chen, and Karin Petrini. "Dynamic audiovisual integration of emotional signals". In: *Cognition and Emotion* 34.5 (2020), pp. 1065–1078.

[228] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. "DECAF: MEG-based multimodal database for decoding affective physiological responses". In: *IEEE Transactions on Affective Computing* 6.3 (2015), pp. 209–222.

[229] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. "Minedojo: Building open-ended embodied agents with internet-scale knowledge". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 18343–18362.

[230] Julia F Christensen, Sebastian B Gaigg, Antoni Gomila, Peter Oke, and Beatriz Calvo-Merino. "Enhancing emotional experiences to dance through music: the role of valence and arousal in the cross-modal bias". In: *Frontiers in human neuroscience* 8 (2014), p. 757.

[231] James Armitage and Tuomas Eerola. "Cross-modal transfer of valence or arousal from music to word targets in affective priming?" In: *Auditory Perception & Cognition* 5.3-4 (2022), pp. 192–210.

[232] Hussain-Abdulah Arjmand, Jesper Hohagen, Bryan Paton, and Nikki S Rickard. "Emotional responses to music: Shifts in frontal brain asymmetry mark periods of musical change". In: *Frontiers in psychology* 8 (2017), p. 2415.

[233] James Z Wang *et al.* "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion". In: *Proceedings of the IEEE* 111.10 (2023), pp. 1236–1286.

[234] Aimee Jeehae Kim. "Differential Effects of Musical Expression of Emotions and Psychological Distress on Subjective Appraisals and Emotional Responses to Music". In: *Behavioral Sciences* 13.6 (2023), p. 491.

[235] American Music Therapy Association. *Standards of practice*. AMTA, 2013.

[236] Katrin Starcke, Johanna Mayr, and Richard von Georgi. "Emotion modulation through music after sadness induction—The iso principle in a controlled experimental study". In: *International journal of environmental research and public health* 18.23 (2021), p. 12486.

[237] Jennifer D Jones. “A comparison of songwriting and lyric analysis techniques to evoke emotional change in a single session with people who are chemically dependent”. In: *Journal of music therapy* 42.2 (2005), pp. 94–110.

[238] Reyhaneh Akbari, Shole Amiri, and Hosseinali Mehrabi. “The effectiveness of music therapy on reducing alexithymia symptoms and improvement of peer relationships”. In: *International Journal of Behavioral Sciences* 14.4 (2021), pp. 178–184.

[239] Orii McDermott, Martin Orrell, and Hanne Mette Ridder. “The development of music in dementia assessment scales (MiDAS)”. In: *Nordic journal of music therapy* 24.3 (2015), pp. 232–251.

[240] Damien K. Ming, Sorawat Sangkaew, Ho Quang Chanh, Pham Thanh Nhat, Sophie Yacoub, Pantelis Georgiou, and Alison H. Holmes. “Continuous physiological monitoring using wearable technology to inform individual management of infectious diseases, public health and outbreak responses”. In: *International Journal of Infectious Diseases* 96 (2020), pp. 648–654. DOI: 10.1016/j.ijid.2020.05.086.

[241] Vaishnavi Bhaltadak, Babaji Ghewade, and Seema Yelne. “A Comprehensive Review on Advancements in Wearable Technologies: Revolutionizing Cardiovascular Medicine”. In: *Cureus* 16.5 (2024). DOI: 10.7759/cureus.12345.

[242] Niels TB Scholte *et al.* “A scoping review on advancements in noninvasive wearable technology for heart failure management”. In: *NPJ Digital Medicine* 7.1 (2024), pp. 1–15. DOI: 10.1038/s41746-023-00925-5.

[243] Badziili Nthubu. “An overview of sensors, design and healthcare challenges in smart homes: future design questions”. In: *Healthcare* 9.10 (2021). DOI: 10.3390/healthcare9101232.

[244] David Pickham *et al.* “Effect of a wearable patient sensor on care delivery for preventing pressure injuries in acutely ill adults: A pragmatic randomized clinical trial (LS-HAPI study)”. In: *International Journal of Nursing Studies* 80 (2018), pp. 12–19. DOI: 10.1016/j.ijnurstu.2017.12.012.

[245] H. Ceren Ates *et al.* "End-to-end design of wearable sensors". In: *Nature Reviews Materials* 7.11 (2022), pp. 887–907. DOI: 10.1038/s41578-022-00460-x.

[246] Chan Wang *et al.* "Artificial intelligence enhanced sensors-enabling technologies to next-generation healthcare and biomedical platform". In: *Bioelectronic Medicine* 9.1 (2023), p. 17. DOI: 10.1186/s42234-023-00118-1.

[247] Shaghayegh Shajari *et al.* "The emergence of AI-based wearable sensors for digital health technology: a review". In: *Sensors* 23.23 (2023), p. 9498. DOI: 10.3390/s23239498.

[248] Tomasz Wasilewski, Wojciech Kamysz, and Jacek Gębicki. "AI-Assisted Detection of Biomarkers by Sensors and Biosensors for Early Diagnosis and Monitoring". In: *Biosensors* 14.7 (2024), p. 356. DOI: 10.3390/bios14070356.

[249] Bonnie Spring *et al.* "Healthy apps: mobile devices for continuous monitoring and intervention". In: *IEEE Pulse* 4.6 (2013), pp. 34–40. DOI: 10.1109/MPUL.2013.2279620.

[250] Muhammad Ali Shiwani *et al.* "Continuous monitoring of health and mobility indicators in patients with cardiovascular disease: A review of recent technologies". In: *Sensors* 23.12 (2023), p. 5752. DOI: 10.3390/s23125752.

[251] Vangelis D. Karalis. "The integration of artificial intelligence into clinical practice". In: *Applied Biosciences* 3.1 (2024), pp. 14–44. DOI: 10.3390/applbiosci3010002.

[252] Talha Iqbal *et al.* "Towards integration of artificial intelligence into medical devices as a real-time recommender system for personalised healthcare: State-of-the-art and future prospects". In: *Health Sciences Review* (2024). DOI: 10.1016/j.hsr.2024.100150.

[253] Mingrui Chen *et al.* "Artificial Intelligence-Based Medical Sensors for Healthcare System". In: *Advanced Sensor Research* 3.3 (2024), p. 2300009. DOI: 10.1002/adsr.202300009.

[254] Alessandro Zompanti, Panaiotis Finamore, Filippo Longo, Simone Grasso, Luca Frasca, Federica Celoro, Marco Santonico, Costanza Cenerini, Ludovica La Monica, Anna Sabatini, *et al.* "Sensor technology advancement enhancing exhaled breath portability: Device

set up and pilot test in the longitudinal study of lung cancer". In: *Sensors and Actuators B: Chemical* 423 (2025), p. 136735.

[255] *World Health Organization*. URL: https://www.who.int/health-topics/diabetes#tab=tab_1.

[256] Sanjay Basu and John S. Yudkin. "Estimation of global insulin use for type 2 diabetes, 2018-30: a microsimulation analysis". In: *Lancet Diabetes Endocrinol* (2018).

[257] Jinli Liu, Ruhai Bai, Zhonglin Chai, Mark E Cooper, Paul Z Zimmet, and Lei Zhang. "Low-and middle-income countries demonstrate rapid growth of type 2 diabetes: an analysis based on Global Burden of Disease 1990–2019 data". In: *Diabetologia* 65.8 (2022), pp. 1339–1352.

[258] Katja Dovc and Tadej Battelino. "Time in range centered diabetes care". In: *Clin Pediatr Endocrinol* 30.1 (2021), pp. 1–10. DOI: 10.1297/cpe.30.1.

[259] Lutz Heinemann. "Continuous Glucose Monitoring (CGM) or Blood Glucose Monitoring (BGM): Interactions and Implications". In: *J Diabetes Sci Technol* 12.4 (2018), pp. 873–879. DOI: 10.1177/1932296818768834.

[260] Thomas Martens *et al.* "Effect of Continuous Glucose Monitoring on Glycemic Control in Patients With Type 2 Diabetes Treated With Basal Insulin: A Randomized Clinical Trial". In: *JAMA* 325.22 (2021), pp. 2262–2272. DOI: 10.1001/jama.2021.7444.

[261] Yoeri M Luijf *et al.* "Accuracy and reliability of continuous glucose monitoring systems: a head-to-head comparison". In: *Diabetes technology therapeutics* 15.8 (2013), pp. 722–727. DOI: 10.1089/dia.2013.0049.

[262] Andrea Facchinetti, Giovanni Sparacino, and Claudio Cobelli. "Modeling the error of continuous glucose monitoring sensor data: critical aspects discussed through simulation studies". In: (2010).

[263] Martina Vettoretti *et al.* "Development of an error model for a factory-calibrated continuous glucose monitoring sensor with 10-day lifetime". In: *Sensors* 19.23 (2019), p. 5320.

[264] Anna Sabatini, Costanza Cenerini, Luca Vollero, and Danilo Pau. "Impact of Interfering Factors on a Glucose Sensor Model". In: *2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE. 2024, pp. 1–6.

[265] Anna Sabatini, Costanza Cenerini, Luca Vollero, and Danilo Pau. "Calibrating Glucose Sensors at the Edge: A Stress Generation Model for Tiny ML Drift Compensation". In: *BioMedInformatics* 4.2 (2024), pp. 1519–1530.

[266] Jan S Krouwer and George S Cembrowski. "A review of standards and statistics used to describe blood glucose monitor performance". In: *Journal of Diabetes Science and Technology* 4.1 (2010), pp. 75–83.

[267] Andrea Facchinetti *et al.* "Modeling the glucose sensor error". In: *IEEE Transactions on Biomedical Engineering* 61.3 (2013), pp. 620–629.

[268] Andrea Facchinetti, Simone Del Favero, Giovanni Sparacino, and Claudio Cobelli. "Model of glucose sensor error components: identification and assessment for new Dexcom G4 generation devices". In: *Medical & biological engineering & computing* 53 (2015), pp. 1259–1269.

[269] Martina Vettoretti, Andrea Facchinetti, Giovanni Sparacino, and Claudio Cobelli. "A model of self-monitoring blood glucose measurement error". In: *Journal of diabetes science and technology* 11.4 (2017), pp. 724–735.

[270] Martina Vettoretti, Simone Del Favero, Giovanni Sparacino, and Andrea Facchinetti. "Modeling the error of factory-calibrated continuous glucose monitoring sensors: application to Dexcom G6 sensor data". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 750–753.

[271] Chengyuan Liu *et al.* "Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal". In: *Sensors* 19.19 (2019), p. 4338.

[272] Andrew DeHennis and Mark Mortellaro. "CGM sensor technology". In: *Glucose Monitoring Devices*. Elsevier, 2020, pp. 111–134.

[273] Hisao Ichijo, Hatsuho Uedaira, Tetsuro Suehiro, Jun'ichi Nagasawa, Aizo Yamauchi, and Noboru Aisaka. "Thermal stability of free and immobilized glucose oxidase studied by activity assay and calorimetry". In: *Agricultural and biological chemistry* 53.3 (1989), pp. 833–834.

[274] EO Odebunmi and SO Owalude. "Kinetic and thermodynamic studies of glucose oxidase catalysed oxidation reaction of glucose". In: (2007).

[275] Chiara Fabris and Marc D Breton. "Modeling the CGM measurement error". In: *Glucose Monitoring Devices*. Elsevier, 2020, pp. 241–253.

[276] Shridhara Alva, Timothy Bailey, Ronald Brazg, Erwin S Budiman, Kristin Castorino, Mark P Christiansen, Gregory Forlenza, Mark Kipnes, David R Liljenquist, and Hanqing Liu. "Accuracy of a 14-day factory-calibrated continuous glucose monitoring system with advanced algorithm in pediatric and adult population with diabetes". In: *Journal of Diabetes Science and Technology* 16.1 (2022), pp. 70–77.

[277] Saroj Kumar Das, Kavya K Nayak, PR Krishnaswamy, Vinay Kumar, and Navakanta Bhat. "Electrochemistry and other emerging technologies for continuous glucose monitoring devices". In: *ECS Sensors Plus* (2022).

[278] Chiara Dalla Man *et al.* "The UVA/PADOVA Type 1 Diabetes Simulator: New Features". In: *Journal of diabetes science and technology* 8.1 (2014), pp. 26–34. DOI: 10.1177/1932296813514502.

[279] Giada Acciaroli, Martina Vettoretti, Andrea Facchinetti, Giovanni Sparacino, and Claudio Cobelli. "Reduction of Blood Glucose Measurements to Calibrate Subcutaneous Glucose Sensors: A Bayesian Multiday Framework". In: *IEEE Transactions on Biomedical Engineering* 65.3 (2017), pp. 587–595.

[280] Soumyabrata Talukder, Souvik Kundu, and Ratnesh Kumar. "Dynamic Calibration of Nonlinear Sensors with Time-Drifts and Delays by Bayesian Inference". In: *arXiv preprint arXiv:2208.13819* (2022).

[281] Jinyu Xie. *Simglucose v0.2.1*. 2018.

[282] *Dropsense C110D*. https://www.dropsens.com/en/pdfs_productos/new_brochures/110-c110.pdf. Accessed: 2023-11-14.

[283] Martina Drecogna, Martina Vettoretti, Simone Del Favero, Andrea Facchinetti, and Giovanni Sparacino. "Data gap modeling in continuous glucose monitoring sensor data". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 4379–4382.

[284] Michele Schiavon, Chiara Dalla Man, Simmi Dube, Michael Slama, Yogish C Kudva, Thomas Peyser, Ananda Basu, Rita Basu, and Claudio Cobelli. "Modeling plasma-to-interstitium glucose kinetics from multitracer plasma and microdialysis data". In: *Diabetes Technology & Therapeutics* 17.11 (2015), pp. 825–831.

[285] Giada Acciaroli, Martina Vettoretti, Andrea Facchinetti, and Giovanni Sparacino. "Calibration of CGM systems". In: *Glucose Monitoring Devices*. Elsevier, 2020, pp. 173–201.

[286] Inc Dexcom. *User guide G7*. https://dexcompdf.s3.us-west-2.amazonaws.com/en-us/G7-CGM-Users-Guide.pdf. Accessed: (03-10-2024). 2023.

[287] Shridhara Alva *et al.* "Accuracy of a 14-day factory-calibrated continuous glucose monitoring system with advanced algorithm in pediatric and adult population with diabetes". In: *Journal of Diabetes Science and Technology* 16.1 (2022), pp. 70–77.

[288] Giada Acciaroli, Martina Vettoretti, Andrea Facchinetti, and Giovanni Sparacino. "Calibration of minimally invasive continuous glucose monitoring sensors: state-of-the-art and current perspectives". In: *Biosensors* 8.1 (2018), p. 24.

[289] *STM32Cube.AI Developer Cloud*. https://stm32ai-cs.st.com/home. Accessed: 2023-4-3.

[290] Md Fahim Rabby, Yuchun Tu, Md Irfan Hossen, In Lee, Abdullah-Al Maida, and Xiaojun Hei. "Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction". In: *BMC medical informatics and decision making* 21.1 (2021), p. 101.

[291] Camilo Mosquera-Lopez and Peter G Jacobs. “Incorporating glucose variability into glucose forecasting accuracy assessment using the new glucose variability impact index and the prediction consistency index: An LSTM case example”. In: *Journal of Diabetes Science and Technology* 16 (2022), pp. 7–18.

[292] Brian A Bogue Jimenez. “Exploring Noninvasive Features for Continuous Glucose Monitoring”. MA thesis. University of Memphis, 2021.

[293] François Chollet *et al. Keras*. https://keras.io. 2015.

[294] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[295] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the properties of neural machine translation: Encoder-decoder approaches”. In: *arXiv preprint arXiv:1409.1259* (2014).

[296] Aaron Voelker, Ivana Kajić, and Chris Eliasmith. “Legendre memory units: Continuous-time representation in recurrent neural networks”. In: *Advances in neural information processing systems* 32 (2019).

[297] Ashutosh Pandey and DeLiang Wang. “TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 6875–6879. DOI: 10.1109/ICASSP.2019.8683634.

[298] Panaiotis Finamore, Simone Scarlata, and Raffaele Antonelli Incalzi. “Breath analysis in respiratory diseases: state-of-the-art and future perspectives”. In: *Expert Review of Molecular Diagnostics* 19.1 (2019), pp. 47–61.

[299] I Horváth, PJ Barnes, S Loukides, PJ Sterk, M Högman, AC Olin, A Amann, B Antus, E Baraldi, A Bikov, *et al.* “A European Respiratory Society technical standard: exhaled biomarkers in lung disease”. In: *European Respiratory Journal* 49.4 (2017).

[300] Roberto F Machado, Daniel Laskowski, Oliver Deffenderfer, Timothy Burch, Shuo Zheng, Peter J Mazzone, Tarek Mekhail, Constance Jennings, James K Stoller, Jacqueline Pyle, *et al.* "Detection of lung cancer by sensor array analyses of exhaled breath". In: *American Journal of Respiratory and Critical Care Medicine* 171.11 (2005), pp. 1286–1291.

[301] Rocco Rocco, Raffaele Antonelli Incalzi, Giorgio Pennazza, Marco Santonico, Claudio Pedone, Simone Bartoli, Chiara Vernile, Elena Frezzotti, Marco Santonico, and Arnaldo D'Amico. "BIONOTE e-nose technology may reduce false positives in lung cancer screening programmes". In: *European Journal of Cardio-Thoracic Surgery* 49.4 (2016), pp. 1112–1117.

[302] Yoav Y Broza, Rom Kremer, Ulrike Tisch, Anna Gevorkyan, Abid Shiban, Lael A Best, and Hossam Haick. "A nanomaterial-based breath test for short-term follow-up after lung tumor resection". In: *Nanomedicine: Nanotechnology, Biology and Medicine* 9.1 (2013), pp. 15–21.

[303] Inbar Nardi-Agmon, Manal Abud-Hawa, Ori Liran, Nitzan Gai-Mor, Maya Ilouze, Amir Onn, Jair Bar, Ofer Golan, Nir Peled, and Hossam Haick. "Exhaled breath analysis for monitoring response to treatment in advanced lung cancer". In: *Journal of Thoracic Oncology* 11.6 (2016), pp. 827–837.

[304] Helin Koç, Julian King, Gerald Teschl, Karl Unterkofler, Susanne Teschl, Pawel Mochalski, Hartmann Hinterhuber, and Anton Amann. "The role of mathematical modeling in VOC analysis using isoprene as a prototypic example". In: *Journal of Breath Research* 5.3 (2011), p. 037102.

[305] Julian King, Alexander Kupferthaler, Karl Unterkofler, Helin Koc, Susanne Teschl, Gerald Teschl, Wolfram Miekisch, Jochen Schubert, Hartmann Hinterhuber, and Anton Amann. "Isoprene and acetone concentration profiles during exercise on an ergometer". In: *Journal of Breath Research* 3.2 (2009), p. 027006.

[306] Sebastien Antoni, Jacques Ferlay, Isabelle Soerjomataram, Ariana Znaor, Ahmedin Jemal, and Freddie Bray. "Bladder cancer incidence and mortality: a global overview and recent trends". In: *European Urology* 71.1 (2017), pp. 96–108.

[307] 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". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249.

[308] S Manship *et al.* "Experiences and Support Needs of People Living With Advanced Cancer: A Systematic Review". In: *Journal of Clinical Oncology* (2024).

[309] Lars Dyrskjøt, Donna E Hansel, Jason A Efstathiou, Margaret A Knowles, Matthew D Galsky, Jeremy Teoh, and Dan Theodorescu. "Bladder cancer". In: *Nature Reviews Disease Primers* 9.1 (2023), p. 58.

[310] Seema Yadav, Anuja Yadava, Saif Hassan, SFA Zaidi, Abbas Zaidi, and Zhichao Luo. "Glucose and tissue oximetry sensors: A review of non-invasive wearable sensors". In: *Biosensors and Bioelectronics* 223 (2023), p. 115065.

[311] IARC. *Global Cancer Observatory: Cancer Tomorrow*. 2023. URL: https://gco.iarc.fr/tomorrow/en.

[312] Arni Kristjansson and H Sigurdsson. "Designing technology for conscious light based skin lesion discrimination". In: *Psychology Research and Behavior Management* 9 (2016), p. 249.

[313] Simon Colton, Geraint A Wiggins, *et al.* "Computational creativity: The final frontier?" In: *Ecai*. Vol. 12. Montpelier. 2012, pp. 21–26.

[314] Fabio Crimaldi and Manuele Leonelli. "AI and the creative realm: A short review of current and future applications". In: *arXiv preprint arXiv:2306.01795* (2023).

[315] Margaret A Boden and Ernest A Edmonds. "What is generative art?" In: *Digital Creativity* 20.1-2 (2009), pp. 21–46.

[316] Tula Giannini and Jonathan P Bowen. "Generative Art and Computational Imagination: Integrating poetry and art". In: *Proceedings of EVA London 2023*. BCS Learning & Development. 2023, pp. 211–219.

[317] Ziv Epstein, Aaron Hertzmann, Laura Herman, Robert Mahari, Morgan R Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, Jessica Fjeld, *et al.* "Art and the science of generative AI: A deeper dive". In: *arXiv preprint arXiv:2306.04141* (2023).

[318] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, *et al.* "Art and the science of generative AI". In: *Science* 380.6650 (2023), pp. 1110–1111.

[319] Joo-Wha Hong and Nathaniel Ming Curran. "Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.2s (2019), pp. 1–16.

[320] Vera Molnar. "Toward aesthetic guidelines for paintings with the aid of a computer". In: *Leonardo* (1975), pp. 185–189.

[321] Paul Cohen. "Harold Cohen and AARON". In: *Ai Magazine* 37.4 (2016), pp. 63–66.

[322] Pamela McCorduck. "Artificial intelligence: An apercu". In: *Daedalus* (1988), pp. 65–83.

[323] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[324] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[325] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. "Inceptionism: Going deeper into neural networks". In: (2015).

[326] John Burns. "The Next Rembrandt". In: *Multimedia Technology Reviews* (2016).

[327] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[328] Jonathan Jones. "A portrait created by AI just sold for $432,000." In: *The Guardian* (2018).

[329] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, *et al.* "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[330] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. "Vqgan-clip: Open domain image generation and editing with natural language guidance". In: *European Conference on Computer Vision*. Springer. 2022, pp. 88–105.

[331] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, *et al.* "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25278–25294.

[332] Matias del Campo and Neil Leach. "Unleashing New Creativities". In: *Architectural Design* 92.3 (2022), pp. 122–135.

[333] Antonio Somaini. "On the photographic status of images produced by generative adversarial networks (GANs)". In: *Philosophy of Photography* 13.1 (2022), pp. 153–164.

[334] Anna Notaro. "State-of-the-art: AI through the (artificial) Artist's Eye". In: *EVA London 2020: Electronic Visualisation and the Arts* (2020), pp. 322–328.

[335] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. "Text-to-image diffusion model in generative ai: A survey". In: *arXiv preprint arXiv:2303.07909* (2023).

[336] Omri Avrahami, Ohad Fried, and Dani Lischinski. "Blended latent diffusion". In: *arXiv preprint arXiv:2206.02779* (2022).

[337] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, *et al.* "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36479–36494.

[338] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[339] A. M. Turing. "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236 (Oct. 1950), pp. 433–460. ISSN: 0026-4423.

[340] Robert M French. “The Turing Test: the first 50 years”. In: *Trends in cognitive sciences* 4.3 (2000), pp. 115–122.

[341] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. “Turing test: 50 years later”. In: *Minds and machines* 10.4 (2000), pp. 463–518.

[342] Graham R Oppy and David L Dowe. “The Turing Test”. In: *Stanford Encyclopedia of Philosophy* 1 (2003), pp. 1–26.

[343] Stevan Harnad. “The Turing Test is not a trick: Turing indistinguishability is a scientific criterion”. In: *ACM SIGART Bulletin* 3.4 (1992), pp. 9–10.

[344] David MW Powers. “The total Turing test and the Loebner prize”. In: *New Methods in Language Processing and Computational Natural Language Learning*. 1998.

[345] Margaret A Boden. “The Turing test and artistic creativity”. In: *Kybernetes* 39.3 (2010), pp. 409–413.

[346] Antonio Daniele, Caroline Di Bernardi Luft, and Nick Bryan-Kinns. ““What Is Human?” A Turing Test for Artistic Creativity”. In: *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*. Springer. 2021, pp. 396–411.

[347] Marian Mazzone and Ahmed Elgammal. “Art, creativity, and the potential of artificial intelligence”. In: *Arts*. Vol. 8. 1. MDPI. 2019, p. 26.

[348] Nils Köbis and Luca D Mossink. “Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry”. In: *Computers in human behavior* 114 (2021), p. 106553.

[349] Regina Schober. “Passing the Turing test? AI generated poetry and posthuman creativity”. In: *Artificial Intelligence and Human Enhancement: Affirmative and Critical Approaches in the Humanities* 21 (2022), p. 151.

[350] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. “AI choreographer: Music conditioned 3D dance generation with AIST++”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13401–13412.

[351] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text-conditional image generation with CLIP latents”. In: *arXiv preprint arXiv:2204.06125* (2022).

[352] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[353] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, *et al.* “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252.

[354] Vivian Liu and Lydia B Chilton. “Design guidelines for prompt engineering text-to-image generative models”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–23.

[355] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. “Seeing is not always believing: Benchmarking human and model perception of ai-generated images”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[356] Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. “As good as a coin toss human detection of ai-generated images, videos, audio, and audiovisual stimuli”. In: *arXiv preprint arXiv:2403.16760* (2024).