ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

COMPUTER SCIENCE AND ENGINEERING

Ciclo 38

Settore Concorsuale: 09/H1 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

Settore Scientifico Disciplinare: ING-INF/05 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

MULTIMODAL AI FOR HUMAN EXPRESSION UNDERSTANDING

Presentata da: Eleonora Mancini

Coordinatore Dottorato

Paola Salomoni

Supervisore

Paolo Torroni

Co-supervisore

Federico Chesani

Esame finale anno 2026

"The most beautiful thing we can experience is the mysterious.
It is the source of all true art and science.
He to whom this emotion is a stranger, who can no longer pause
to wonder and stand rapt in awe, is as good as dead:
his eyes are closed."
Albert Einstein

Abstract

Human expression is inherently multimodal, spanning language, sound, movement, and other perceptual forms that together convey meaning. Each modality carries distinct yet complementary information, and understanding arises from their integration rather than from any single channel. Yet most Artificial Intelligence (AI) systems still process these signals in isolation, failing to capture an important part of the complexity and richness of human expression.

This thesis investigates multimodal AI across three domains where such integration is essential: political discourse analysis, clinical assessment, and music information retrieval. It addresses three key challenges that limit progress in multimodal AI: obtaining ethically sound and representative multimodal data, designing effective architectures for multimodal fusion, and developing techniques to provide model explanations that can be trusted and understood by domain experts.

In political discourse analysis, we establish multimodal argument mining by developing for the first time large-scale resources that combine text and speech, filling the gap left open by the absence of audio in existing datasets. We show that prosodic and acoustic cues enhance argument detection. Moreover, we release open toolkits to foster reproducible research in this field.

In clinical assessment, we define guidelines for responsible dataset creation and curation, and perform trimodal fusion of linguistic, acoustic, and neurophysiological signals for depression detection, achieving state-of-the-art results. Moreover,

using Parkinson's detection from speech as a case study, we analyse current post-hoc explanation techniques, showing that they fail to provide clinically actionable insights. This reveals a gap between algorithmic decisions and clinical understanding.

In music information retrieval, we address the opposite data imbalance found in argument mining: while musical audio is abundant, lyrics are scarce, fragmented, or restricted by copyright. This asymmetry hinders multimodal approaches that rely on paired text-audio data. To overcome this limitation, we introduce WEALY, a reproducible pipeline that learns lyrics-aware representations directly from audio, enabling lyrics-based reasoning without textual transcriptions.

Lastly, building on some of our findings, we propose a methodology for perceptually grounded interpretability that can generalise across multimodal domains. We develop listenable explanations that enable experts to hear which acoustic properties drive model decisions, showing that explanations can be both perceptually accessible and faithful to model reasoning.

This thesis demonstrates that comprehensive and ethically grounded multimodal data, principled fusion strategies and perceptually informed interpretability are fundamental in building AI systems that can represent human expression in all its complexity.

Keywords: Multimodal AI, Natural Language Processing, Speech Processing, Multimodal Fusion, Explainable AI, Dataset Creation, AI Ethics, Argument Mining, Clinical Decision Support, Music Information Retrieval

Contents

Abstract **II**

1 Introduction **1**

1.1 Problem Definition 2

1.1.1 The Multimodal Imperative 2

1.1.2 The Fundamental Challenges of Multimodal AI 7

1.2 Research Methodology 13

1.2.1 Open Science and Reproducibility 13

1.2.2 Responsible Data Development and Usage 14

1.2.3 Computational Resource Management and Accessibility . . 15

1.3 Research Questions 15

1.3.1 Cross-Domain 15

1.3.2 Domain-Specific 16

1.4 Structure of the Dissertation 17

2 Background **19**

2.1 Representation 19

2.1.1 Text Representation 20

2.1.2 Speech and Audio Representation 26

2.1.3 EEG Signal Processing and Representation 32

2.2 Multimodal Deep Learning 34

2.2.1 Challenges of Multimodal Deep Learning 35

3 Multimodal Argument Mining **37**

3.1 Introduction 37

3.1.1 Challenges 39

3.1.2 Our Contribution 41

3.1.3 Chapter Structure 42

3.2 Related Work 44

3.2.1 Foundations 44
3.2.2 Prior Multimodal Argument Mining Research 45
3.2.3 Multimodal Deep Learning for Multimodal Argument Mining 46
3.2.4 Standardization and Tools 47
3.2.5 Research Gaps 48
3.3 Multimodal Argument Mining: A Case Study in Political Debates 49
3.3.1 Datasets and Corpus Development 49
3.3.2 Methodology and Experimental Design 52
3.3.3 Results and Analysis 55
3.3.4 Discussion and Findings 59
3.4 Multimodal Fallacy Classification in Political Debates 61
3.4.1 Domain Description 61
3.4.2 Structural Foundation 62
3.4.3 Corpus Development 64
3.4.4 Corpus Statistics 65
3.4.5 Methodology and Experimental Design 68
3.4.6 Results and Analysis 69
3.4.7 Discussion and Findings 71
3.5 MAMKit: A Comprehensive Multimodal Argument Mining Toolkit 72
3.5.1 MAMKit 73
3.5.2 Experiments 82
3.5.3 Results 82
3.5.4 Discussion and Findings 83
3.6 Overview of MM-ArgFallacy2025 on Multimodal Argumentative Fallacy Detection and Classification in Political Debates 85
3.6.1 Problem Formulation 85
3.6.2 Data 86
3.6.3 Overview of the Systems and Results 89
3.6.4 Main Findings 91
3.6.5 Discussion 93
3.7 Conclusion 96

4 Multimodal Clinical Decision Support 99
4.1 Introduction 99
4.1.1 Challenges 100
4.1.2 Our Contribution 104

4.1.3 Chapter Structure . . . 105
4.2 Related Work . . . 106
4.2.1 Ethical Concerns and Dataset Documentation . . . 106
4.2.2 Multimodal Depression Detection from Clinical Interviews . 108
4.2.3 Explainability in Speech-Based Clinical Detection . . . 109
4.3 Promoting the Responsible Development of Speech Datasets for Mental Health and Neurological Disorders Research . . . 111
4.3.1 A Case Study on DAIC-WoZ Depression Research . . . 112
4.3.2 Desiderata . . . 113
4.3.3 Methodology . . . 122
4.3.4 Analysis . . . 125
4.3.5 Findings . . . 137
4.4 TRI-DEP: A Trimodal Comparative Framework for Depression Detection Using Speech, Text, and EEG . . . 139
4.4.1 Methodology . . . 140
4.4.2 Experimental Setup . . . 145
4.4.3 Results . . . 146
4.4.4 Findings . . . 150
4.5 Investigating the Effectiveness of Explainability Methods in Parkinson's Detection from Speech . . . 152
4.5.1 Methodology . . . 153
4.5.2 Experiments . . . 155
4.5.3 Results . . . 156
4.5.4 Findings . . . 159
4.6 Conclusion . . . 161

5 Multimodal Musical Version Identification 165
5.1 Introduction . . . 165
5.1.1 Challenges . . . 167
5.1.2 Our Contribution . . . 169
5.1.3 Chapter Structure . . . 170
5.2 Related Work . . . 171
5.2.1 Musical Version Identification . . . 171
5.2.2 Lyrics-Based Music Retrieval . . . 171
5.2.3 Speech Foundation Models for Music Information Retrieval 173
5.2.4 Multimodal Fusion in Music Information Retrieval . . . 173

5.3 Leveraging Whisper Embeddings for Audio-based Lyrics Matching 176
5.3.1 Method . . . 176
5.3.2 Experimental Setup . . . 180
5.3.3 Results . . . 182
5.3.4 Findings . . . 184
5.4 Conclusion . . . 186

6 Roadmap Towards Explainable Multimodal AI 189
6.1 Introduction . . . 189
6.1.1 Our Contribution . . . 192
6.1.2 Chapter Structure . . . 193
6.2 Related Work . . . 193
6.3 LMAC-TD: Producing Time Domain Explanations for Audio Classifiers . . . 196
6.3.1 Methodology . . . 197
6.3.2 Experiments . . . 199
6.3.3 Findings . . . 204
6.4 Conclusion . . . 205

7 Conclusion 207
7.1 Summary of Contributions . . . 209
7.2 Discussion and Future Work . . . 211

Acknowledgments 214

List of Figures 218

List of Tables 221

List of Abbreviations 223

Bibliography 228

Appendix 276

CHAPTER 1

Introduction

Human expression is inherently multimodal, weaving together diverse streams of information that transcend what any single channel could convey alone. When we engage in meaningful conversations, assess clinical dialogues, or experience musical performances, we naturally integrate information from multiple channels, processing not just what is said (lexical choices, syntactic structures, and semantic content), but how it is said (prosodic features, vocal quality, and paralinguistic cues), when it is said, and with what emotional and contextual undertones. In clinical contexts, this extends to neurophysiological signals such as Electroencephalography (EEG), where event-related potentials, spectral power distributions, and neural synchronisation patterns reveal cognitive processing states and responses to stimuli. This inherent multimodality of human expression represents both the richness of human communication and a critical challenge for AI systems seeking to understand and interact with humans in natural settings. Despite this reality, most computational approaches to language understanding, clinical assessment, and information retrieval have remained unimodal, focusing primarily on either audio or textual content while ignoring the wealth of information conveyed by

the other modalities. This disconnect between how humans naturally communicate and how machines process information has created a fundamental gap in AI capabilities, limiting the effectiveness of systems designed to support human decision-making in critical domains. This thesis addresses this multimodal gap through systematic research across three domains in which the integration of different modalities is not merely beneficial but essential for meaningful progress: political discourse analysis, clinical decision support, and Music Information Retrieval (MIR). Each domain presents unique challenges and opportunities that collectively illuminate the broader principles and technical requirements for effective multimodal AI systems.

1.1 Problem Definition

1.1.1 The Multimodal Imperative

The motivation for multimodal approaches extends beyond technical curiosity to practical necessity. In political discourse, the persuasiveness of arguments often depends as much on delivery as on logical content [144]. For instance, hesitant speech can undermine sound reasoning, while confident delivery can mask logical fallacies [145]. Citizens and analysts could use tools to assess both what politicians say and how they say it, to make informed democratic decisions. The prosodic features of political speech, including vocal emphasis, pausing patterns, and emotional undertones, convey crucial information about speaker credibility, argument strength, and persuasive intent that is often overlooked in textual analysis alone.

In clinical settings, speech patterns provide crucial diagnostic information that complements verbal content, creating opportunities for more comprehensive and objective patient assessment [269, 184]. Changes in vocal characteristics, speech rhythm, and prosodic features can indicate early stages of neurological conditions like Parkinson's disease, mental health states such as depression and anxiety, and cognitive decline associated with dementia [252, 283, 33]. This information, invisible to text-only analysis, is nevertheless essential for comprehensive patient assessment. Moreover, integrating speech analysis with other physiological signals creates even richer diagnostic possibilities. For instance, combining speech patterns with EEG data enables multimodal analysis that captures not only what patients say and how they say it, but also associated neurophysiological activity that correlates with cognitive processing and affective states. This trimodal ap-

proach integrates semantic content extracted from automatic speech transcription, acoustic features from the speech signal itself, and brain activity captured through EEG recordings. Together, these modalities provide unprecedented insight into mental health conditions, offering objective biomarkers that can complement subjective clinical assessments and potentially detect early indicators of neurological or psychiatric conditions before they manifest in traditional diagnostic criteria.

In musical information retrieval, the relationship between audio and textual content creates complex semantic associations crucial for tasks like Musical Version Identification (MVI). A musical version may preserve lyrical content while dramatically altering musical arrangement, tempo, key, and instrumentation, or maintain musical elements like melody and harmonic structure while modifying, translating, or entirely replacing textual content [370]. Some versions transform both musical and lyrical elements while maintaining enough similarity to be recognisable as versions of the original composition. Understanding these relationships requires integrated analysis of both modalities, as neither audio-only nor text-only approaches can capture the full complexity of musical version relationships that humans intuitively recognise. Furthermore, the legal and commercial implications of musical similarity assessment demand systems capable of distinguishing between melodic, harmonic, rhythmic, and lyrical forms of similarity [378], as copyright law often considers these elements separate intellectual properties with distinct ownership and licensing requirements.

Building on this motivation, the following sections examine how the multimodal imperative manifests in practice across three domains: political discourse, clinical assessment, and musical expression, each exemplifying a different rationale for multimodal integration.

Political Discourse: The Democratic Necessity

Political communication exemplifies the essential nature of multimodal analysis. When citizens evaluate political arguments, they naturally integrate multiple information streams: the logical structure of arguments, the emotional appeal conveyed through vocal emphasis, the confidence or uncertainty revealed through speech patterns, and the authenticity suggested by the alignment between content and delivery.

Research in argumentation theory has long recognised that the persuasiveness of political discourse depends not only on logical soundness but also on rhetorical

delivery, such as how speakers use prosodic features like pace, tone, and emphasis to enhance or undermine their arguments [24]. This multimodal reality becomes particularly critical when analysing argumentative fallacies in political discourse. Consider the difference between a politician delivering a personal attack (ad hominem) with confident, authoritative vocal patterns versus the same content delivered with hesitant, uncertain prosody [143, 98, 144]. While the logical fallacy remains the same, the persuasive impact, and thus the democratic consequences, can be dramatically different. Similarly, emotional appeals (appeal to pity) often rely heavily on prosodic features like slower rhythm, particular vocal qualities, and strategic pauses to maximise their emotional impact.

Text-only analysis of such arguments fundamentally misrepresents their actual effect on audiences. The democratic implications of this gap are profound [177]. In an era of increasing concern about misinformation, political polarisation, and the quality of public discourse, citizens need tools that can help them critically evaluate political communication as they actually experience it, through the full multimodal context of political speech. Media literacy education, fact-checking systems, and democratic transparency tools that ignore the prosodic dimensions of political communication systematically fail to address how political persuasion actually works [362, 30, 151]. Furthermore, the development of such tools faces unique challenges in the political domain. While the theoretical necessity for multimodal analysis is clear, the practical challenges are substantial. Most existing argument mining datasets and resources are text-only, reflecting years of research that have focused on the more tractable problem of textual analysis. Creating comprehensive multimodal political discourse datasets requires not only collecting audio-visual content but also developing sophisticated alignment techniques to connect speech segments with textual annotations, establishing new annotation frameworks that account for prosodic features, and creating evaluation methodologies that can assess the contribution of different modalities to argumentative analysis.

Clinical Assessment: The Diagnostic Imperative

The clinical domain presents perhaps the most compelling case for multimodal analysis, where integrating speech and text can literally be a matter of life and death. Clinicians naturally process not only what patients say but how they say it, recognising that linguistic and paralinguistic cues together convey critical

diagnostic information [62]. Vocal characteristics, speech rhythm, pause patterns, and prosodic features can reveal early signs of neurological conditions such as Parkinson's disease, cognitive decline associated with dementia, and mental health disorders such as depression [252, 283, 33].

Research in clinical speech analysis confirms that speech carries rich multimodal biomarkers often detectable before patients or even experienced clinicians notice overt symptoms. Subtle changes in voice quality, speaking rate, or articulatory precision can indicate the onset of neurodegenerative disease months or years before traditional diagnostic criteria are met. Similarly, acoustic and linguistic features reflecting emotional prosody can provide objective, complementary measures of mood states, enhancing the precision and consistency of clinical assessment [85].

However, realising the potential of multimodal AI in healthcare entails profound ethical and practical challenges [138]. Patient data is inherently sensitive, requiring rigorous protection of privacy, informed consent, and compliance with regulatory frameworks. These constraints limit large-scale data collection, creating a paradox: the field where multimodal AI could have the greatest clinical benefit is also the one where data access is most restricted.

Beyond data availability, multimodal clinical systems must satisfy exceptionally high standards of reliability, interpretability, and fairness, as their outputs directly inform diagnosis and treatment. Clinicians must understand not only what a model predicts but also why it makes specific recommendations, which features drive its conclusions, and how confident it is in them [4]. This interpretability requirement is particularly demanding for multimodal architectures, where interactions between speech, text, and other signals produce complex decision boundaries that can be difficult to translate into clinically meaningful insights. Furthermore, speech patterns vary across demographic and linguistic groups, making fairness in multimodal models essential to ensure equitable assessment and avoid perpetuating disparities [366].

Musical Expression: The Creative and Legal Imperative

Music represents the most inherently multimodal form of human artistic expression, where the relationship between audio and textual content creates complex semantic associations that define artistic meaning and legal ownership. The challenge of MVI, determining whether two musical pieces are versions of the same

underlying composition, illustrates the essential nature of multimodal analysis in creative domains. When an artist performs a new version of an existing song, they may preserve lyrical content while dramatically altering musical arrangement, tempo, key, and instrumentation. Alternatively, musical versions may maintain musical elements like melody and harmonic structure while modifying or translating lyrical content. Some musical versions transform both musical and lyrical elements while maintaining enough similarity to be recognisable as versions of the original [370]. This complexity means that effective MVI requires integrated analysis of both musical and lyrical content, as neither modality alone provides sufficient information for reliable identification.

The stakes of accurate MVI extend beyond technical interest to fundamental questions of creative rights and legal ownership. In the music industry, the distinction between musical composition—typically associated with melody, harmony, and lyrics—and sound recording, referring to the specific performance and production, creates complex copyright relationships in which different parties may own different aspects of the same musical work [338]. MVI systems must therefore distinguish between legitimate artistic reinterpretation and copyright infringement, requiring a nuanced understanding of how different modalities contribute to musical similarity.

Yet developing such systems poses unique challenges. While vast amounts of musical audio exist in digital form, corresponding textual content, particularly lyrics and version annotations, remains limited due to copyright restrictions and sparse documentation [114]. In addition, Automatic Speech Recognition (ASR) approaches are ill-suited for musical audio, as they struggle with vocal-instrumental overlap, expressive singing styles, and production effects that distort linguistic clarity [154].

At the same time, practical MIR applications must scale to collections containing millions of songs [25], demanding multimodal architectures that remain efficient and robust under limited supervision. This combination of abundant raw data, scarce annotations, and high scalability requirements makes music a distinctive and challenging testing ground for multimodal AI.

These constraints exemplify, in the creative sphere, the broader data challenges that pervade multimodal AI, which are explored in detail in the following sections.

1.1.2 The Fundamental Challenges of Multimodal AI

While the necessity of multimodal approaches varies across domains, the technical and methodological challenges of building effective multimodal AI systems exhibit remarkable consistency. Through systematic analysis across political discourse, clinical assessment, and music information retrieval, three fundamental challenges emerge that must be addressed in any successful multimodal AI system.

The Data Challenge: Domain-Specific Obstacles to Multimodal Integration

The first fundamental challenge lies in obtaining appropriate multimodal data for system development and evaluation. This challenge manifests differently across domains but consistently represents the primary barrier to multimodal AI deployment. Unlike unimodal systems that can leverage large-scale text or audio datasets independently, multimodal systems require precisely aligned, high-quality paired data that captures meaningful relationships between speech and text modalities.

In political discourse analysis, the data challenge primarily concerns **alignment and recovery**. Extensive textual resources exist for political speech analysis, including debate transcripts, speech manuscripts, and argument-annotated corpora developed over years of text-focused research [94, 103]. However, the corresponding audio content is often missing, misaligned, or of insufficient quality for analysis. Creating comprehensive multimodal political discourse datasets requires not only locating original audio sources but also developing sophisticated alignment techniques to connect speech segments with existing textual annotations. This process is complicated by the challenges of segmenting continuous political speeches into sentence-level units, handling variations between prepared remarks and actual delivery, and managing the acoustic complexities of live political events, including audience reactions, multiple speakers, and variable audio quality. The political domain also faces unique challenges related to **temporal coverage and representation**. Political discourse evolves over time, reflecting changing social concerns, communication technologies, and rhetorical styles. Multimodal political datasets must capture sufficient temporal breadth to support robust model development while ensuring balanced representation across different political perspectives, speaking styles, and demographic characteristics of speakers.

In clinical settings, the data challenge is dominated by **ethical constraints and regulatory requirements**. While clinical encounters naturally generate rich multimodal data through patient-clinician interactions, the collection and

use of such data for research purposes must navigate complex ethical landscapes, including patient privacy protection, informed consent procedures, and institutional review board oversight. These requirements are not merely procedural obstacles but reflect fundamental ethical principles about patient autonomy and data protection that must be respected by any responsible clinical AI development. The clinical domain faces additional challenges related to **data representativeness and bias prevention**. Clinical speech patterns vary significantly across demographic groups, and historical biases in healthcare delivery mean that existing clinical datasets may not adequately represent diverse patient populations [366]. Developing equitable clinical multimodal AI systems requires careful attention to demographic balance, linguistic diversity, and socioeconomic representation in training datasets. Moreover, clinical multimodal datasets must meet exceptionally high standards for **data quality and annotation accuracy**. Clinical assessments have direct implications for patient care, meaning that training datasets must include not only high-quality audio-text pairs but also reliable clinical ground truth labels that reflect actual diagnostic outcomes. This requirement often necessitates longitudinal data collection that tracks patient outcomes over extended periods, further complicating data collection efforts [138].

In musical information retrieval, the data challenge reflects **copyright restrictions and annotation scarcity**. While enormous quantities of musical content exist in digital form, copyright protections limit research access to commercial recordings. Even when audio content is accessible through academic licensing agreements, the associated textual metadata, particularly lyrical content and version relationships, is often sparse or entirely absent. The music domain also faces challenges related to **scalability and annotation cost**. Practical musical information retrieval systems must operate on collections containing millions of songs, but manual annotation of musical lyrics is prohibitively expensive at this scale. This creates a fundamental tension between the need for large-scale datasets and the practical limitations of human annotation efforts [114].

The Fusion Challenge: Integrating Complementary and Competing Information

The second fundamental challenge lies in developing effective architectures for combining information across multiple modalities. This challenge extends beyond simple feature concatenation to encompass the complex problem of learning

meaningful cross-modal representations that can handle both complementary and contradictory information streams.

Complementary Integration — In many cases, different modalities provide complementary information that enhances overall system understanding. In political discourse, textual content provides explicit logical structure while prosodic features convey implicit information about speaker confidence, emotional state, and rhetorical intent. In clinical settings, patient responses provide semantic content while speech patterns offer objective biomarkers that may not be reflected in verbal reports, and EEG signals can provide neurophysiological indicators that complement both speech and text analysis. In musical applications, lyrical content provides semantic meaning while audio features capture melodic, harmonic, and rhythmic relationships. Effective fusion architectures must be capable of learning these complementary relationships without requiring explicit specification of how modalities interact. This requires, for example, developing attention mechanisms that can identify relevant cross-modal connections, representation learning techniques that can capture shared semantic spaces across modalities, and architectural designs that allow for flexible modality weighting based on task requirements and input characteristics [167].

Contradictory Resolution — Multimodal systems must also handle cases where different modalities provide contradictory information. In political analysis, speakers may deliver confident-sounding rhetoric while their prosodic patterns reveal uncertainty or deception. In clinical settings, patients may verbally report feeling well while their speech patterns indicate underlying health concerns, or their EEG signals may reveal neurological activity inconsistent with their self-reported mental state. In musical applications, musical versions may share lyrical content while differing dramatically in musical style, or maintain musical elements while translating or modifying textual content. Resolving these contradictions requires sophisticated fusion architectures that can weigh modality contributions dynamically based on task context, input quality, and learned reliability patterns [63]. This is particularly challenging because the optimal resolution strategy may vary not only across domains but also across different instances within the same domain.

Temporal Alignment — An additional complexity in multimodal fusion concerns temporal alignment across different modalities with varying temporal char-

acteristics. While text represents discrete symbolic information with clear segmentation boundaries, speech unfolds continuously over time with complex temporal dynamics, and physiological signals like EEG operate at different temporal scales with millisecond-level precision. Effective fusion architectures must account for these temporal differences while learning meaningful alignments between elements across modalities. This temporal alignment challenge is particularly acute in scenarios involving multiple continuous signals [337]. Political debates include interruptions, overlapping speech, and off-script remarks that complicate alignment with prepared textual content. Clinical interviews involve natural conversational dynamics that may not align neatly with structured assessment frameworks, while physiological signals may exhibit patterns that span different temporal windows than verbal exchanges. An additional complication arises when EEG data are not recorded simultaneously with speech and text modalities, requiring sophisticated alignment techniques that can account for temporal offsets and establish correspondences across non-concurrent recording sessions. Musical performances may include instrumental sections, repeated verses, and improvised elements that require flexible alignment strategies across audio and textual content.

Modality-Specific Preprocessing — Different modalities also require different preprocessing approaches that can affect fusion performance. Text processing typically involves tokenization, normalization, and embedding generation that produces discrete, symbolic representations. Audio processing may involve feature extraction at multiple temporal scales, from frame-level acoustic features to utterance-level prosodic summaries. Physiological signals like EEG require specialized preprocessing including artifact removal, frequency band analysis, and temporal windowing. Effective fusion architectures must account for these preprocessing differences while learning meaningful cross-modal associations [17].

The Explainability Challenge: Making Multimodal Decisions Interpretable

The third fundamental challenge concerns developing interpretable multimodal AI systems that can provide meaningful explanations for their decisions to human users. This challenge is particularly critical because multimodal systems operate in high-stakes domains where users need to understand not only what systems recommend but why specific decisions are made and how confident systems are in their analyses.

The Audio Explainability Gap — While techniques for explaining text-based AI decisions are relatively well-developed, including attention visualisation, feature-importance analysis, and example-based explanations, the explainability of audio models remains largely unresolved [128]. The conceptualisation and vocabulary that humans adopt to describe textual features are closely related to the language model representation of text. On the contrary, the human descriptions of audio features are distant from machine representations of audio and speech. This divergence reflects both a perceptual and a representational gap: humans experience sound through temporal dynamics, spectral characteristics, and intensity variations, yet translating these perceptual impressions into precise linguistic terms is inherently difficult. This asymmetry between auditory perception and verbal description poses significant challenges for generating explanations that are both meaningful to non-expert users and faithful to model reasoning.

Audio explainability is further complicated by the continuous nature of speech. While textual explanations can highlight contributory words or phrases, speech explanations must identify relevant time intervals and acoustic properties within continuous signals. Effective methods, therefore, need to bridge perception and representation by mapping temporal and spectral information into interpretable formats, a task made difficult by our limited capacity to verbalise auditory experiences with the specificity achieved for other modalities.

Domain-Specific Explanation Requirements — Different application domains require different types of explanations tailored to specific user needs and decision contexts. In political discourse analysis, citizens need explanations that can help them understand how both argument content and delivery characteristics contribute to assessments of argument strength or fallacy identification. These explanations must be accessible to general audiences without specialised knowledge of speech processing or argumentation theory. In clinical settings, explanations must serve multiple stakeholder groups with different expertise levels and information needs. Clinicians require detailed technical explanations that identify specific speech biomarkers and their clinical significance. Patients need accessible explanations that help them understand assessment results and their implications for treatment. Researchers need comprehensive explanations that can support hypothesis generation and biomarker discovery. In musical applications, explanations must serve diverse stakeholders, including legal professionals evaluating copyright claims, music industry professionals making licensing decisions, and re-

searchers studying musical similarity. Each stakeholder group requires different levels of technical detail and different types of evidence to support decision-making processes [141, 160].

Cross-Modal Attribution — Multimodal explainability also requires developing techniques for attributing decisions to specific modalities and cross-modal interactions [358]. Users need to understand whether specific decisions are driven primarily by textual content, speech characteristics, or interactions between modalities. This attribution challenge is complicated by the complex non-linear interactions that effective multimodal fusion architectures create between input modalities. Furthermore, multimodal explanations must account for temporal dynamics in both speech and text processing. Users may need to understand not only which features are important but also how the importance of different features changes over time during the processing of extended speech samples [309]. Finally, multimodal explainability systems require robust evaluation methodologies that can assess the quality and utility of explanations across different user groups and application contexts. This evaluation challenge is complicated by the subjective nature of explanation quality and the difficulty of establishing ground truth for explanation correctness [228].

The challenges just outlined related to multimodal data, fusion, and explainability fundamentally shape this thesis's methodological approach. Recent multimodal foundation models like GPT-5.2 [235] and Gemini 3 [96] appear to obviate specialised research with their native multimodality. However, fundamental limitations make them unsuitable for the research goals of this thesis.

It has been observed that the most famous multimodal foundation models demonstrate systematic modality bias, over-relying on text while marginalising acoustic information [356, 41]. Moreover, they support only common modalities (text, audio, images, video), excluding neurophysiological signals like EEG required for clinical applications. Additionally, for discriminative tasks—which encompass all applications in this thesis—their generation-optimised architectures are computationally inefficient and architecturally misaligned with classification objectives.

This thesis, therefore, develops transparent, specialised architectures with explicit control over multimodal representation, fusion, and interpretation. By addressing cross-modal integration, data-efficient learning, and interpretable decision-making in data-scarce domains, this research establishes principles that

can inform next-generation multimodal foundation models.

1.2 Research Methodology

Throughout this research, we have been guided by three core methodological principles that ensure both technical rigour and practical applicability across diverse application domains. These principles reflect not only technical considerations but also ethical commitments to responsible AI development and community benefit.

1.2.1 Open Science and Reproducibility

The first guiding principle concerns commitment to open science practices that enable community validation, extension, and impact. This commitment is particularly crucial in multimodal AI research, where dataset access has historically been a significant barrier to progress and where the complexity of multimodal systems makes reproducibility challenging.

Dataset Publication — Part of this research required developing new multimodal datasets. All multimodal datasets developed in this research are publicly available with comprehensive documentation including collection methodologies, annotation protocols, data format specifications, and baseline experimental results. This includes MM-USElecDeb60To16 (MM-USED) for political discourse analysis, MM-USElecDeb60To16-fallacy (MM-USED-fallacy) for fallacy classification. These datasets are published with persistent identifiers and long-term preservation commitments to ensure continued community access.

Code and Model Release — All experimental code, model implementations, and training procedures are published in version-controlled repositories with comprehensive documentation and dependency management. This includes not only final model implementations but also data preprocessing pipelines, evaluation frameworks, and experimental analysis scripts. Model checkpoints and trained parameters are also released where feasible to enable direct comparison and extension.

Experimental Transparency — All experimental protocols, hyperparameter selections, and evaluation procedures are documented in detail with sufficient information to enable exact replication. This includes reporting negative results, ablation studies, and sensitivity analyses that provide a comprehensive under-

standing of model behaviour and limitations.

Community Engagement — This research contributes to community standards through participation in shared tasks (such as MM-ArgFallacy2025), organisation of research workshops, and collaborative development efforts. This engagement helps establish common evaluation frameworks and promotes coordination across research groups working on related problems.

1.2.2 Responsible Data Development and Usage

The second guiding principle concerns ethical responsibility in data collection, curation, and usage practices. Each application domain presents unique ethical considerations that require careful attention to privacy protection, bias mitigation, and fair representation.

Ethical Framework Development — In the clinical domain, we developed comprehensive ethical frameworks for the responsible development of speech datasets for mental health and neurological disorder research. This framework addresses data collection, storage, sharing, and usage practices through consultation with domain experts, institutional review boards, and affected communities to ensure alignment with both technical requirements and ethical obligations.

Domain-Specific Ethical Considerations — In clinical applications, we conducted a systematic analysis of how existing datasets have been collected and curated, identifying ethical gaps and proposing guidelines for responsible clinical speech data development. For political discourse datasets, we addressed ethical limitations and concerns in the collection and dissemination of public debate data, ensuring balanced representation while respecting privacy considerations. In musical applications, we navigated copyright considerations and intellectual property constraints while developing methods that enable research advancement without infringing on artists' rights.

Long-term Stewardship — Data publication includes long-term stewardship plans that address ongoing maintenance, version control, error correction, and community governance. These plans ensure that datasets remain valuable community resources beyond the lifetime of the project they were developed for.

1.2.3 Computational Resource Management and Accessibility

The third guiding principle addresses the significant computational requirements of multimodal AI research while promoting accessibility and sustainability. Multimodal research demands extensive computational resources not only for model training and inference but also for data construction, alignment, and preprocessing pipelines. Creating multimodal datasets requires aligning speech with text transcriptions, synchronising audio with video, or matching physiological signals with behavioural data. These processes are computationally intensive and time-consuming, and combined with modelling demands, create potential barriers to participation for researchers with limited access to high-performance computing infrastructure.

Resource Optimisation — All model development includes systematic optimisation efforts to minimise computational requirements while maintaining performance. This includes architecture design choices that balance model capacity with computational efficiency, training procedures that minimize resource usage, and inference optimizations that enable deployment on diverse hardware platforms.

Infrastructure Partnerships — The research leverages strategic partnerships with high-performance computing centers including Compute Canada, ISCRA-C with Cineca, and other national computing facilities. These partnerships enable large-scale experiments while demonstrating sustainable approaches to resource-intensive AI research.

1.3 Research Questions

This thesis addresses multiple research questions, some cross-domain, other domain-specific.

1.3.1 Cross-Domain

RQ1: Multimodal Necessity — How do multimodal approaches to human expression analysis compare to unimodal methods across critical application domains, and what evidence supports the necessity of integrating multiple communication channels?

RQ2: Cross-Domain Challenges — What are the fundamental challenges that must be addressed to develop effective multimodal AI systems across diverse ap-

plication domains, and how do these challenges manifest differently in political discourse, clinical assessment, and musical analysis?

RQ3: Systematic Solutions — How can systematic approaches to data collection, fusion architectures, and explainability techniques advance the state of multimodal AI across critical human-centred applications?

1.3.2 Domain-Specific

Multimodal Argument Mining

- **RQ4: Data Alignment Challenge** — How can existing text-based argument mining resources be effectively aligned with audio content to create comprehensive multimodal political discourse datasets?
- **RQ5: Multimodal Performance** — To what extent do multimodal approaches improve argument structure detection and fallacy classification compared to text-only methods in political discourse analysis?

Multimodal Clinical Decision Support

- **RQ6: Ethical Framework Development** — What ethical frameworks are necessary for responsible development and use of multimodal clinical datasets, particularly for mental health and neurological disorder research?
- **RQ7: Trimodal Clinical Assessment** — How does the integration of speech, text, and EEG signals improve clinical assessment accuracy compared to unimodal and bimodal approaches, particularly for depression detection?
- **RQ8: Clinical Speech Biomarkers** — What speech-based biomarkers can be identified and validated for neurological conditions like Parkinson's disease, and how can these biomarkers be made interpretable to clinicians?

Multimodal Musical Information Retrieval

- **RQ9: Large-Scale Lyrics Data Challenge** — How can meaningful lyrical representations be obtained for large-scale music collections when copyright restrictions and data scarcity prevent access to comprehensive lyrics datasets?

- **RQ10: Multimodal Performance** — How effectively can simple fusion approaches integrate lyrics-aware and audio content representations to capture complementary aspects of musical similarity for version identification tasks?

1.4 Structure of the Dissertation

This thesis comprises six main chapters exploring multimodal AI through applications in argument mining, clinical decision support, and music information retrieval, ultimately proposing a unified cross-domain explainability methodology. Given the interdisciplinary scope spanning multimodal argument mining, clinical decision support, music information retrieval, and Explainable Artificial Intelligence (XAI), each chapter includes a dedicated *related work* section situating the research within its specific domain.

Chapter 2, *Background*, provides the technical underpinnings of data representation and multimodal deep learning. Section 2.1 investigates the transformation of diverse communication modalities into computational representations. Section 2.2 introduces multimodal deep learning, examining early fusion, late fusion, and hybrid fusion paradigms. The chapter closes with fundamental challenges in multimodal learning.

Chapter 3, *Multimodal Argument Mining*, develops multimodal approaches to argument mining in political debates. Section 3.2 reviews related work. Section 3.3 introduces MM-USED, the largest multimodal argument mining dataset at release, presenting architectural evaluations across core tasks and comparing traditional with advanced multimodal techniques. Section 3.4 extends to fallacy classification with MM-USED-fallacy, the first multimodal fallacy corpus. Section 3.5 presents MAMKit, a standardising PyTorch toolkit. Section 3.6 describes the first shared task on multimodal fallacy analysis. Section 3.7 synthesises contributions and future directions.

Chapter 4, *Multimodal Clinical Decision Support*, investigates multimodal approaches for mental health and neurological assessment. Section 4.2 reviews computational approaches, data ethics, fusion strategies, and XAI in healthcare. Section 4.3 establishes an ethical framework for speech data collection and curation, critically reviewing 36 papers. Section 4.4 presents TRI-DEP, a trimodal depression detection study comparing features, architectures, and fusion strategies. Section 4.5 benchmarks post-hoc XAI approaches for Parkinson's disease detection.

Section 4.6 discusses clinical implications and interpretability requirements.

Chapter 5, *Multimodal Music Information Retrieval*, explores multimodal integration of lyrics and audio content. Section 5.2 reviews related work. Section 5.3 presents WEALY, a method to extract lyrics-aware representations from raw audio using Whisper embeddings and contrastive learning. Multimodal fusion experiments validate complementarity, achieving state-of-the-art performance. Section 5.4 synthesizes contributions and demonstrates foundation model transfer capabilities.

Chapter 6, *Roadmap Towards Explainable Multimodal AI*, addresses explaining decisions in speech-based multimodal systems, establishing *listenability* as a fundamental property. Section 6.2 reviews XAI methods. Section 6.3 presents LMAC-TD, a method that produces time-domain interpretations enabling domain experts to hear which acoustic properties drive decisions. Section 6.4 establishes listenability as a measurable property extensible across all thesis domains.

Chapter 7 integrates contributions, revisits research questions, and proposes future directions for multimodal AI research.

CHAPTER 2

Background

The integration of multiple modalities, such as speech/audio, text, and neurophysiological signals, requires computational representations that capture the unique characteristics of each modality while enabling meaningful cross-modal interactions. This section establishes the first layer of the multimodal deep learning framework by examining how different forms of human expression are transformed into machine-readable representations. Effective representation is a prerequisite for multimodal AI, as it determines how information can be aligned, fused, and interpreted across heterogeneous modalities.

The final section of this chapter builds on the challenges introduced in Chapter 1, examining how multimodal representations can be integrated and jointly modelled, and discussing the key challenges in learning from multiple modalities.

2.1 Representation

Over the last decades, two fundamental paradigms have emerged for transforming raw modalities into computational features:

- **Handcrafted features** rely on domain expertise to design interpretable descriptors grounded in linguistic theory, signal processing, or neuroscience principles. These features are data-efficient and interpretable but fixed by design, unable to adapt to specific tasks. Moreover, feature engineering is labour-intensive and requires substantial domain expertise for each application, limiting the scalability of machine learning systems [22].
- **Representation learning** discovers features directly from data through optimisation, reducing dependence on manual feature engineering [22, 164]. Methods range from static embeddings learned from distributional statistics to contextual representations from self-supervised pre-training to end-to-end learning within task-specific neural architectures. While effective at capturing complex patterns, these approaches require more training data and produce less interpretable features than handcrafted approaches.

The following subsections detail how these paradigms manifest for each modality, emphasising the specific methods employed in this thesis. For speech, we first present time-frequency representations (spectrograms, mel-spectrograms) that serve as foundational signal transforms for both handcrafted feature extraction and learned representations. For text, we additionally cover classical approaches (sparse vector representations and n-gram language models) that predate neural approaches, providing important historical context and serving as interpretable baselines in our experiments.

2.1.1 Text Representation

Text processing focuses on extracting semantic meaning, structural relationships, and contextual information from discrete symbolic input. Unlike continuous signals, text provides explicit symbolic structure through discrete tokens that can be processed using fundamentally different computational approaches.

For text, classical approaches rely on sparse vector representations of documents, n-gram language models for sequential structure, and handcrafted linguistic features encoding domain knowledge. Neural approaches learn dense, distributed representations through optimisation on large corpora. Each captures different aspects of textual information and serves distinct roles in multimodal systems.

Sparse Vector Representations

Sparse vector methods represent documents in high-dimensional spaces where each dimension corresponds to a vocabulary term.

Bag-of-Words and TF-IDF — The *bag-of-words (BoW)* model encodes each document as a vector of term counts, disregarding word order or syntax [285]:

$$\mathbf{d} = [c(w_1), c(w_2), \ldots, c(w_{|V|})], \tag{2.1}$$

where $c(w_i)$ is the count of term w_i in the document, and $|V|$ is the vocabulary size.

To reduce bias from frequent but uninformative words, the *Term Frequency-Inverse Document Frequency (TF-IDF)* weighting scheme [129] scales each term by its relative importance across documents:

$$\text{TF-IDF}(t_i, d) = \frac{\text{tf}(t_i, d)}{\max_k \text{tf}(t_k, d)} \cdot \log \frac{|D|}{|\{d \in D : t_i \in d\}|}, \tag{2.2}$$

where $\text{tf}(t_i, d)$ is the term frequency in document d, $|D|$ is the total number of documents, and the denominator counts documents containing t_i. TF-IDF became the cornerstone of traditional information retrieval systems and baseline text classifiers.

In this thesis, TF–IDF vectors are used as text features in some experiments for the argument-mining tasks (Section 3.3) to build classical SVM baselines, and as part of the transcription-based baselines for WEALY (Section 5.3) to evaluate lyrics similarity from Whisper-generated transcriptions.

N-gram Language Models

To capture local word order, *n-gram models* represent text as sequences of n consecutive tokens [296, 48]. The probability of a word sequence $w_1, \ldots, w_n$ is approximated as:

$$P(w_1, w_2, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-k+1}, \ldots, w_{i-1}), \tag{2.3}$$

where k is the context window (e.g., $k = 2$ for bigrams, $k = 3$ for trigrams). Despite their simplicity, n-gram models remain foundational, serving as the basis for early speech recognition and machine translation systems.

Handcrafted Linguistic Features

Beyond raw statistics, traditional Natural Language Processing (NLP) systems often relied on explicitly engineered linguistic features that capture structure and meaning:

- Part-of-speech tag distributions and syntactic parse-tree patterns [202];
- Named-entity counts and semantic type ratios [295];
- Sentiment and affective lexicon scores using resources such as LIWC [324].

Limitations — Classical methods, though interpretable, exhibit critical limitations. Sparse vector representations ignore word order and long-range dependencies, yielding high-dimensional vectors that fail to generalise to unseen words. N-gram models suffer from exponential growth in parameters with context size and cannot capture long-range dependencies. Handcrafted features require extensive linguistic expertise and preprocessing. These shortcomings motivated the shift toward dense distributed neural representations that discover patterns directly from data.

Learned Representations

Representation learning overcomes the limitations of sparse statistical methods by discovering dense, continuous embeddings from data. The foundation underlying neural text representations is *language modelling*. Language modelling seeks to estimate the probability distribution of word sequences, thereby capturing both syntactic and semantic regularities of language [23]. Formally, a language model computes:

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \prod_{i=2}^{n} P(w_i \mid w_1, \ldots, w_{i-1}), \tag{2.4}$$

where each conditional term models the likelihood of a word given its preceding context. Performance is commonly evaluated via *perplexity*. Perplexity quantifies how well a probabilistic model predicts unseen data. Between two models, the one that assigns higher probability to the test set, i.e., provides a tighter fit to its distribution, is considered superior [131]. For a test set $W = w_1, w_2, ...w_N$, the perplexity is the probability of the test set, normalised by the number of words:

$$\text{Perplexity} = P(w_1, w_2, ...w_N)^{-1/N}, \tag{2.5}$$

with lower values indicating better predictive accuracy.

Building on this language modelling foundation, neural text representation learning has evolved through three major stages: distributional word embeddings that learn from co-occurrence patterns, self-supervised language models that produce contextualised representations, and specialised models adapted for specific tasks.

A. Distributional Word Embeddings — The distributional hypothesis posits that words appearing in similar contexts have similar meanings [107]. Distributional word embedding models operationalise this principle by learning dense, fixed-length vector representations from word co-occurrence patterns in large corpora.

Word2Vec [209] learns embeddings through shallow neural prediction tasks. The Skip-gram objective predicts context words from a target word, while the Continuous Bag-of-Words (CBOW) variant predicts target words from their context.

GloVe [250] (Global Vectors) combines global matrix factorisation with local context learning, factorising word co-occurrence matrices to produce embeddings that exhibit linear semantic regularities (e.g., king $-$ man $+$ woman $\approx$ queen). GloVe embeddings serve as input to sequence-based architectures in argument mining tasks (Section 3.3).

Limitations and motivation for self-supervised language models. Static embeddings assign each word a single fixed vector, failing to capture polysemy (multiple meanings) or context-dependent semantics. The word "bank" receives identical representations whether referring to a financial institution or a riverbank. Furthermore, these models cannot handle out-of-vocabulary words and provide no mechanism for modeling compositional semantics at the sentence or document level. These limitations motivated the development of self-supervised language models that produce different embeddings for the same word depending on its surrounding context.

B. Self-Supervised Language Models — Self-supervised language models produce *contextualized representations* where each token's embedding depends on its surrounding context, learned through language modeling objectives on vast amounts of unlabeled text.

The transformer architecture [345] enabled this breakthrough by replacing re-

current computation with self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.6}$$

where Q, K, and V denote query, key, and value matrices, and d_k is the dimensionality of K. This architecture enables parallel processing of sequences and effective modelling of long-range dependencies.

Modern self-supervised language models fall into three architectural families based on their use of the transformer encoder and decoder.

Encoder-Based Models (Bidirectional Understanding). Encoder-only models produce bidirectional contextual representations ideal for understanding-oriented tasks.

BERT [70] achieves this through a masked language modelling (MLM) objective, training the model to infer missing tokens from their bidirectional context.

RoBERTa [179] further optimises BERT through dynamic masking (changing which tokens are masked across epochs), longer training, and training on larger corpora, demonstrating that careful optimisation of the pretraining procedure significantly improves performance.

Building upon BERT's foundation, several variants refine or extend its bidirectional contextual modelling:

ALBERT [159] introduces parameter-sharing across transformer layers and factorised embedding parameterisation, significantly reducing model size while maintaining performance.

DeBERTa [110] (Decoding-enhanced BERT with disentangled attention) improves upon BERT by using disentangled attention mechanisms that separately encode content and position information, and an enhanced mask decoder for better position-aware predictions.

ModernBERT [354] incorporates architectural improvements, including rotary positional embeddings, GeGLU activations, and optimized attention mechanisms, demonstrating that careful architecture design can improve BERT-style models.

MacBERT [61] refines BERT's masked language modelling by replacing the standard random masking with a *masked correction* strategy, where tokens are substituted by plausible alternatives predicted by an auxiliary model, thereby narrowing the pretraining-fine-tuning gap.

XLNet [368] introduces permutation language modelling, which maximises the expected likelihood over all possible word-order permutations, enabling bidirectional context learning without explicit masking and combining the strengths of autoregressive and autoencoding pretraining.

MPNet [274] integrates masked and permuted language modelling by predicting masked tokens conditioned on both left and right context in their natural order, achieving better position encoding and sentence-order consistency than prior transformer encoders.

In this work, BERT and RoBERTa serve as text encoders for multimodal architectures in argument-mining experiments (Sections 3.3–3.5). In TRI-DEP (Section 4.4), we employ BERT, Chinese BERT Base [70], MacBERT, XLNet, and MPNet Multilingual to encode Mandarin clinical transcripts for multimodal fusion with speech and EEG signals. Participating teams in the MM-ArgFallacy2025 shared task (Section 3.6) employ BERT, RoBERTa, ALBERT, DeBERTa, and ModernBERT.

Decoder-Based Models (Autoregressive Generation). Decoder-only architectures generate text autoregressively, predicting each token from prior context. They employ causal masking to prevent access to future tokens:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}). \tag{2.7}$$

GPT models [266] pioneered this approach at scale. The progression from GPT-1 to GPT-4 demonstrates how scaling model size and training data enables emergent capabilities such as in-context learning and complex reasoning.

LLaMA [334] and similar models show that training on high-quality, curated corpora can yield strong performance even with smaller parameter counts compared to the largest commercial models.

DeepSeek [66] demonstrates efficient scaling of decoder-only models through mixture-of-experts architectures and optimized training procedures.

Encoder-Decoder Models (Sequence-to-Sequence). Encoder-decoder architectures combine bidirectional encoding with autoregressive decoding, excelling in sequence-to-sequence tasks such as translation and summarization.

T5 [267] unifies all NLP tasks as text-to-text transformations, framing problems from classification to generation under a single framework where both inputs and outputs are text strings.

Instruction-Following Models. Recent instruction-tuned models adapt pre-trained LLMs to follow natural-language commands through Reinforcement Learning from Human Feedback (RLHF) [239]. In this framework, a reward model is trained to estimate human preference between alternative outputs, and the base model is optimised to produce responses that maximise this learned reward. Instruction-following systems such as **ChatGPT** [239], **Claude** [12], and **Gemini** [7] employ RLHF and large-scale instruction tuning to better align model behaviour with human expectations, enabling more controllable and useful interactions.

C. Specialised Sentence and Domain Models. Domain-tuned or task-specific models adapt general-purpose LLMs to specialised applications. **Sentence-BERT (SBERT)** [274] refines BERT to generate semantically meaningful sentence embeddings by aggregating token-level representations through a pooling operation. This design enables robust performance in tasks such as semantic similarity, clustering, and information retrieval.

In this work, SBERT is used both in the argument-mining tasks (Section 3.4) to model sentence-level semantics, and in the WEALY baselines (Section 5.3) to encode Whisper-generated lyrics transcriptions for comparison with our audio-based embeddings.

In summary, neural text representation has evolved from static distributional embeddings to transformer-based self-supervised models that produce rich contextualised representations, fundamentally redefining how textual information is encoded. These learned representations capture syntactic and semantic nuances that handcrafted features cannot, enabling state-of-the-art performance across diverse NLP tasks. However, they typically require substantial data and computational resources, and produce representations whose internal workings are less transparent than traditional linguistic features, motivating ongoing research into data-efficient and interpretable text representation learning.

2.1.2 Speech and Audio Representation

Speech signals encode information at multiple levels beyond lexical content, including prosody, rhythm, and voice quality. Unlike discrete symbolic text, speech is a continuous acoustic waveform shaped by articulatory dynamics, speaker identity, and recording conditions.

For speech and audio, the traditional feature-based paradigm manifests as signal processing techniques that extract spectral descriptors (e.g., MFCCs, spectral centroid) and prosodic features (e.g., fundamental frequency, intensity, rhythm) from time-frequency representations. The neural paradigm manifests as self-supervised models trained on large unlabeled audio corpora, supervised encoder-decoder models trained on labelled speech-text pairs, and contrastive audio-language models. Each approach captures complementary aspects of acoustic information and serves distinct roles in multimodal systems.

This section details these approaches and indicates where each is employed across argument mining (Sections 3.3–3.5), clinical assessment (Section 4.4), music-lyrics modeling (Section 5.3), and explainability research (Sections 4.5, 6.3).

Time-Frequency Representations

Time-frequency representations jointly preserve temporal and spectral information, serving as both intermediate features and direct inputs to neural models [236, 262].

Spectrograms — The *Short-Time Fourier Transform* (STFT) provides a time–frequency representation of a signal by computing local Fourier transforms over short, overlapping windows. In the continuous-time case, the STFT of a signal $x(t)$ with window $w(t)$ is defined as

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j\omega t}\, dt, \tag{2.8}$$

where $w(\tau)$ denotes the window function, commonly a Hann window or Gaussian window centred around zero.

The discrete counterpart, applied to sampled data, computes the transform frame by frame:

$$X(m, \omega) = \sum_{m=0}^{N-1} x[n]\, w[n-m]\, e^{-j\omega n}, \tag{2.9}$$

where $x[n]$ is the discrete-time signal, $w[\cdot]$ is a window function of length N, m denotes the time-frame index (i.e., the position of the window), and ω is the angular frequency variable. The window slides along the signal, and for each position m, the local frequency content of $x[n]$ is estimated.

The *spectrogram* represents the power spectral density of a signal, visualising

how energy is distributed over time and frequency. It is obtained as the squared magnitude of the STFT:

$$\text{spectrogram}\{x(t)\}(\tau, \omega) \equiv |X(\tau, \omega)|^2, \tag{2.10}$$

or, in the discrete case,

$$S(n, k) = |X(n, k)|^2, \tag{2.11}$$

where $S(n, k)$ denotes the energy at time frame n and frequency bin k.

Spectrograms are widely used in audio processing as they reveal how spectral content evolves over time. However, they discard phase information, providing only a partial view of the original signal.

Mel and Log-Mel Spectrograms — To approximate human auditory frequency resolution, the spectrogram is typically passed through a mel filter bank [349], producing a mel-scaled representation in which frequency bands are spaced according to perceptual sensitivity. Applying logarithmic compression yields the *log-mel spectrogram*, the standard input representation for modern ASR and self-supervised audio models [355].

Across all studied domains, audio features serve as front-end inputs to the corresponding models. In the multimodal argument-mining tasks (Sections 3.3-3.4), we employ MFCCs and related spectral descriptors computed from mel filterbanks, providing low-level representations of prosodic and timbral cues. In TRI-DEP, WEALY and in the Parkinson's XAI study, although representations are derived from self-supervised models such as wav2vec 2.0, HuBERT, and WavLM, the raw waveform is first transformed into a time-frequency representation (e.g., STFT or mel spectrogram) before encoding. In LMAC-TD (Section 6.3), spectrograms constitute the classifier input domain from which time-domain explanations are generated.

Handcrafted Acoustic Features

Traditional speech processing relies on engineered features designed to capture specific acoustic properties through established signal processing techniques.

Spectral Features — Among the most widely used spectral representations are the Mel-Frequency Cepstral Coefficientss (MFCCs) [65], obtained by applying a Discrete Cosine Transform (DCT) to the logarithm of mel filterbank energies. The

mel scale itself is a perceptual frequency scale that approximates the non-linear resolution of human hearing [349].

We employ spectral representations across both the multimodal argument-mining (Sections 3.3-3.4) and clinical pipelines (Section 4.4). In argument mining, we extract 40 MFCCs per frame alongside spectral centroid, roll-off, bandwidth, spectral contrast, and chroma features, which capture complementary timbral and harmonic information while implicitly encoding prosodic correlates such as loudness and emphasis. In TRI-DEP, we employ only 40 MFCCs as the spectral component, complemented by explicitly defined prosodic features.

Prosodic Features — Prosodic descriptors capture suprasegmental aspects of speech that convey affective, cognitive, and clinical information beyond phonetic content [325, 83, 292]. Prosodic features include pitch, loudness, rhythm, and temporal organisation. Here, we present representative descriptors relevant to our studies.

Fundamental frequency (F0) represents the rate of vibration of the vocal folds and corresponds perceptually to pitch. It is inversely related to the period of the vocal fold vibration cycle. Faster vibrations produce higher perceived pitch, while slower vibrations yield lower pitch.

Intensity (RMS energy) reflects short-term acoustic power of the speech signal and correlates with the perceptual notion of loudness [262, 292]. It is commonly estimated as the root-mean-square amplitude of the waveform within short analysis windows. This measure is sensitive to articulation effort, vocal stress, and emotional expressiveness, making it a relevant descriptor for both affective and clinical-state analysis.

Speech timing and rhythm capture aspects such as speaking rate, fluency, and expressiveness. Speaking rate is defined as the number of syllables per second of active speech, excluding pauses, while pause rate and phonation time are computed from voiced/unvoiced segmentation. Together, these descriptors indicate fluency and vocal effort.

The complete prosodic set in TRI-DEP includes six handcrafted features — energy, F_0, RMS, pause rate, phonation time, and speech rate — complementing 40 MFCCs to form a 46-dimensional acoustic baseline.

Limitations — Handcrafted features are interpretable and lightweight but may miss task-specific abstractions and long-range dependencies, and can be sensitive to speaker/channel conditions [292]. These limitations motivated the development

of learned representations.

Learned Representations

Representation learning for speech takes three forms based on training supervision: (a) self-supervised models learn from unlabeled audio through pretext tasks such as masked prediction and contrastive learning; (b) supervised models are trained on labelled speech-text pairs for tasks like automatic speech recognition; (c) contrastive paired models align audio and text representations through audio-caption pairs. Together, these approaches provide complementary learned representations for speech processing.

A. Self-Supervised Speech Models — Self-supervised models learn contextual acoustic embeddings from raw waveforms using pretext objectives that uncover phonetic and linguistic structure without labels.

wav2vec 2.0 [11] introduces a framework that maps raw audio to latent speech representations using a convolutional encoder followed by a Transformer context network. The model masks span the latent space and learns to distinguish the true quantised target representation from distractors via a contrastive loss. The quantisation step, learned jointly with the encoder, provides discrete speech units that enable effective pretraining from unlabeled speech.

XLSR-53 [57] extends wav2vec 2.0 to multilingual speech representation learning by pretraining on 53 languages. This cross-lingual approach enables the model to learn language-agnostic acoustic representations that transfer effectively across diverse languages, making it particularly valuable for low-resource language settings and multilingual applications.

HuBERT [117] extends this paradigm by replacing explicit quantisation with iterative masked prediction over cluster-based pseudo-labels. These labels, obtained via offline k-means clustering of acoustic features (e.g., MFCCs), serve as hidden units guiding the model to predict masked segments. Through repeated re-clustering and re-training, HuBERT learns hierarchical acoustic and linguistic abstractions. Language-specific variants, such as Chinese HuBERT Large [326], apply this architecture to monolingual corpora for improved performance on specific languages.

DistilHuBERT [39] applies knowledge distillation to HuBERT, producing a compressed model that retains most of the performance while significantly reducing computational requirements.

WavLM [46] further advances masked prediction by integrating denoising modelling and structural improvements to robustly learn universal speech representations. It augments the HuBERT architecture with gated relative position bias to enhance temporal modelling. It exposes the model to noisy, multi-speaker mixtures during pretraining, thereby improving its ability to handle diverse and overlapping speech conditions. Trained on 94k hours of mixed-domain audio, WavLM achieves state-of-the-art performance across content, speaker, and paralinguistic tasks on the SUPERB benchmark [336].

Across this thesis, we employ these self-supervised models in three domains: wav2vec 2.0, HuBERT, and WavLM for multimodal argument mining (Sections 3.3–3.5); XLSR-53 and Chinese HuBERT Large for clinical depression assessment in TRI-DEP (Section 4.4); and Parkinson's-tuned HuBERT and WavLM for explainability research (Section 4.5). Participating teams in the MM-ArgFallacy2025 shared task (Section 3.6) employ wav2vec 2.0, WavLM, and DistilHuBERT.

B. Supervised Encoder-Decoder Models — Whisper [265] is a large-scale encoder-decoder ASR model trained on 680k hours of multilingual speech-text pairs. Its encoder consumes log-mel spectrograms, and the decoder autoregressively predicts text tokens while generating intermediate hidden states that capture rich acoustic-linguistic context. While not self-supervised, Whisper's learned representations transfer effectively to downstream tasks.

In our music-lyrics study (Section 5.3), we extract these decoder hidden states — termed *lyrics-aware Whisper latents* — as fixed-length audio representations that encode lyrical semantics directly from song audio without explicit transcription. Whisper is also employed by participating teams in the MM-ArgFallacy2025 shared task (Section 3.6) for audio encoding.

C. Contrastive Audio-Language Models — CLAP [363] (Contrastive Language-Audio Pretraining) aligns audio and language representations through dual encoders trained with a contrastive objective on paired audio-text data. Inspired by CLIP [264] in computer vision, CLAP projects audio and text embeddings into a shared multimodal space, enabling zero-shot classification by measuring cosine similarity between embeddings of audio samples and natural language prompts. While trained on paired data, the audio-text associations are unlabeled (no explicit classification labels), making this a form of contrastive paired learning rather than fully supervised learning.

CLAP representations are integrated in argument mining experiments (Sections 3.3, 3.4).

2.1.3 EEG Signal Processing and Representation

EEG measures the brain's electrical activity through scalp electrodes, offering millisecond-level temporal resolution of neural processes. Unlike speech and text, which capture communicative behaviour, EEG directly reflects neurophysiological activity underlying cognition and affect. Its high temporal resolution, multi-channel spatial structure, and multi-scale spectral content require specialised preprocessing and representation methods.

For EEG, the handcrafted paradigm manifests as neuroscience-guided feature extraction that computes descriptors in time, frequency, and spatial domains. The representation learning paradigm manifests as self-supervised foundation models pretrained on large-scale EEG datasets that learn general-purpose neural embeddings through masked prediction over multi-channel recordings. Each approach captures complementary aspects of neurophysiological information and serves distinct roles in clinical applications.

This section details these approaches as employed in our clinical depression assessment study (Section 4.4), where we compare handcrafted descriptors against foundation model embeddings.

Handcrafted EEG Features

Preprocessing and Denoising — Raw EEG signals are inherently low-amplitude and highly susceptible to contamination from ocular, muscular, and environmental sources. The recorded signal at electrode i and time t can be modelled as the sum of the true neural activity and an additive noise term [286, 230]:

$$x_i(t) = s_i(t) + n_i(t), \tag{2.12}$$

where $s_i(t)$ denotes the underlying cerebral activity and $n_i(t)$ includes both random noise and structured artefacts such as eye blinks, muscle movements, or power-line interference.

Preprocessing typically includes band-pass filtering (e.g., 0.5–50 Hz) and a notch filter at the power-line frequency, followed by re-referencing (e.g., average or linked-mastoid reference). In more advanced setups, Independent Component Analysis (ICA) can be employed to identify and remove artefact components [146].

Filtering can equivalently be described in the frequency domain, where the signal's spectrum is multiplied by the desired filter response before transforming back to the time domain [236, 286]. These general preprocessing operations form the foundation of the EEG pipeline adopted in Section 4.4, where filtering and re-referencing are applied to prepare the signals for feature extraction and foundation-model embedding.

Feature Extraction — After preprocessing, handcrafted features are computed in the time, frequency, and spatial domains [286]. Time-domain measures include simple statistics and Hjorth parameters [113]. Spectral features are typically based on power spectral density and canonical band powers (delta, theta, alpha, beta, gamma) [314, 229]. Spatial/functional relations can be summarised with coherence or phase-based metrics [158, 311].

In TRI-DEP, we use a compact set of handcrafted EEG descriptors as baselines (Section 4.4), including statistical features (e.g., mean, variance), spectral power across canonical frequency bands, and entropy-based measures of signal complexity.

Learned Representations

Large-scale EEG foundation models learn general-purpose neural embeddings via self-supervised pre-training, improving cross-dataset robustness and reducing reliance on manual feature design.

Self-Supervised Models — **LaBraM (Large Brain Model)** [125] is among the first large-scale foundation models for EEG representation learning. It tokenises EEG recordings into fixed-length channel-time patches and employs a Transformer trained with a masked prediction objective over vector-quantised neural tokens. This design allows the model to capture both local temporal patterns and global spatial dependencies while remaining robust to heterogeneous electrode layouts. For methodological details, including the neural spectrum tokeniser and pretraining objective, we refer the reader to [125]. The pretrained LaBraM model introduced in that work is used to extract EEG representations in our clinical study (Section 4.4).

CBraMod (Criss-Cross Brain Foundation Model) [352] extends the patch-based paradigm by explicitly modelling the spatial and temporal dependencies of EEG signals through a criss-cross Transformer architecture. It combines spatial and temporal attention mechanisms, asymmetric positional encoding, and

joint time-frequency patch representations to learn context-aware neural embeddings. Further architectural details and training procedures can be found in [352]. The pretrained CBraMod model is likewise employed to extract EEG representations in our clinical study (Section 4.4).

Discussion — Both LaBraM and CBraMod follow the patch-based representation paradigm but operate at distinct abstraction levels. LaBraM applies patching directly to raw EEG segments to unify input representations across datasets, whereas CBraMod constructs latent patch embeddings that integrate temporal and spatial context via dedicated attention pathways. Together, they exemplify complementary strategies for large-scale EEG pretraining, both of which are leveraged in this thesis for clinical applications.

Strengths and Limitations — Foundation models provide strong transfer and reduce manual feature engineering, but require large, diverse corpora and substantial compute; their embeddings are also less directly interpretable than classical features.

2.2 Multimodal Deep Learning

Multimodal Deep Learning (MMDL) is gaining significant attention across diverse AI fields, including natural language processing, computer vision, healthcare, autonomous systems, and human-computer interaction [268]. The success of deep learning has catalysed progress on increasingly complex machine-learning problems involving multiple data modalities. In multimodal settings, data from different sensors or perceptual channels capture complementary aspects of the same phenomenon, enabling richer and more robust representations than unimodal systems can achieve.

A key advantage of deep architectures lies in their ability to learn hierarchical representations automatically, both within each modality and in shared fused spaces, without requiring handcrafted features or extensive preprocessing. End-to-end multimodal training also scales effectively with data size and the number of modalities [268].

Traditionally, multimodal fusion has followed two main paradigms [17]: early fusion, which combines features from different modalities before modality-specific processing, and late fusion, which integrates independent unimodal decisions through voting or averaging. However, these approaches might struggle to cap-

ture complex cross-modal dependencies and temporal interactions. To overcome these limitations, modern neural architectures employ cross-modal attention mechanisms and multimodal transformers that enable each modality to attend to relevant information from the others. An increasingly effective alternative is intermediate fusion, which integrates learned representations across multiple levels of the network, offering finer control over when and how modalities interact.

2.2.1 Challenges of Multimodal Deep Learning

MMDL research focuses on developing neural architectures that can effectively integrate diverse modalities by managing both intra-modal and inter-modal dynamics [17]. Managing these dynamics gives rise to several fundamental challenges that any effective MMDL model must address. These challenges manifest consistently across the political discourse, clinical assessment, and musical domains examined in this work, though with domain-specific variations. Specifically, an effective MMDL model must address [17]:

- **Representation**: Learning machine-interpretable representations for each modality is essential for ensuring compatibility with machine-learning models. The heterogeneity of multimodal data makes this challenging, as different modalities require fundamentally different encoding approaches. In political discourse, this requires encoding textual argument structures and speech prosodic patterns in ways that maximize interpretability for argumentation analysis. In clinical settings, this involves representing semantic content from patient utterances, acoustic biomarkers from speech (voice quality, speaking rate, pitch variability), and neurophysiological patterns from EEG (event-related potentials, spectral power distributions) in formats relevant for diagnostic assessment. In MIR, this requires encoding audio features, lyrical content, and symbolic music representations in ways that maximize interpretability and task relevance.

- **Translation**: This involves transforming data from one modality to another, where relationships between modalities are often open-ended or subjective. In this thesis, we do not address this challenge.

- **Alignment**: Recognising relationships across modalities is critical for effective multimodal integration. In political discourse, this involves aligning speech segments with textual argument annotations and managing temporal

dynamics of debates with interruptions and overlapping speech. In clinical settings, alignment requires synchronising speech patterns with clinical assessment frameworks and connecting patient utterances with corresponding EEG activity, which is particularly challenging when data are recorded non-concurrently. In MIR, this could involve aligning lyrics to specific audio segments or synchronizing symbolic notation with corresponding audio excerpts, thereby enhancing interpretability through cross-modal connections.

- **Fusion**: Aggregating information from multiple modalities is essential for tasks involving classification or decision-making, where different modalities may have varying predictive power and noise characteristics. In political discourse analysis, fusion integrates logical argument structure with prosodic delivery characteristics to assess argument strength and identify fallacies. In clinical decision support, fusion combines semantic content, acoustic biomarkers, and neurophysiological indicators to produce comprehensive diagnostic assessments. In MIR, by integrating insights from lyrics, audio, and symbolic representations, systems can achieve greater robustness and accuracy in version identification and similarity assessment.

- **Co-Learning**: This refers to the transfer of knowledge across modalities, enabling models to generalise more effectively, particularly when one modality has limited resources. In this thesis, we do not address this challenge.

CHAPTER 3

Multimodal Argument Mining

3.1 Introduction

Argument Mining (AM) is a rapidly evolving area within NLP that aims to automatically extract arguments and their structural relations from natural language sources [178]. Its ultimate goal is to computationally model how humans construct, present, and interconnect arguments in discourse by identifying components such as claims and premises, classifying their roles, and uncovering the relationships that link them. Emerging at the intersection of computational linguistics, AI, and argumentation theory, AM addresses the growing need to automatically process the vast amount of argumentative content generated in digital communication.

Performing AM typically involves a set of interconnected tasks: component detection (identifying text spans containing arguments), classification (distinguishing between claims, premises, and non-argumentative content), link and relation prediction (identifying the nature of argumentative connections such as support or attack), and stance detection (determining the position taken toward a

topic) [163]. Each task poses specific challenges, from handling implicit argumentation to accounting for contextual and domain-specific variability.

Among the domains explored for AM, including legal documents, scientific literature, social media, and educational texts, the political domain is particularly compelling. Political discourse is inherently argumentative: politicians seek to persuade, defend positions, attack opponents, and justify policy choices. Consequently, political texts and speeches are rich in diverse argumentative patterns. Moreover, political AM enables socially relevant applications such as detecting logical fallacies [94], persuasiveness degree [37], verifying the truthfulness of claims [225, 150], analyzing coherence and consistency in candidates' reasoning [35, 177], and summarizing political positions [347].

Beyond its academic interest, political AM has clear societal relevance. In a context of information overload and growing polarisation, automated systems for analysing political discourse can support democratic processes by helping citizens interpret arguments, assisting fact-checkers in verifying claims, and enabling researchers to trace the evolution of political communication. Debates and speeches also provide a controlled yet naturalistic setting where argumentative discourse unfolds in real time, making them ideal benchmarks for advancing AM methodologies.

Despite these advantages, most AM research remains text-centric, treating debates and speeches as purely linguistic artefacts. This approach overlooks key aspects of political communication, which is inherently multimodal: beyond words, speakers use prosody, tone, rhythm, and pauses to enhance persuasion. Recent studies have underscored the importance of paralinguistic features in argumentative discourse across various domains [143, 98], as well as in related areas such as fake news detection [121].

The audio modality conveys cues unavailable in transcripts — vocal stress, pitch variation, emotional tone, and speaking rate — that reveal intent and influence perceived argument strength. These signals can aid in identifying argumentative components and assessing persuasive force [177, 348, 255]. Theoretical foundations for integrating audio stem from rhetoric, sociolinguistics, and cognitive science: classical rhetoric recognised delivery (actio) as central to persuasion; sociolinguistic research links prosody to social and semantic meaning; and cognitive studies show that listeners integrate multiple cues when processing persuasive messages [69, 301, 346, 216]. In political contexts, such acoustic strategies help

speakers convey authority, emotion, and credibility — key factors in argumentative effectiveness.

These insights have led to the emergence of Multimodal Argument Mining (MAM), which extends traditional AM by modelling argumentative discourse as an inherently multimodal phenomenon. Drawing on advances in multimodal machine learning, computational linguistics, and argumentation theory, MAM aims to capture how linguistic and paralinguistic features jointly shape argumentation in natural communication.

The rise of MAM aligns with recent progress in multimodal deep learning (MMDL), which has produced powerful representation and fusion mechanisms, including transfer-learning-based architectures [332], attention models for cross-modal alignment [171, 99], and transformer-based systems integrating textual and acoustic information [337]. At the same time, advances in pre-trained models for text and speech have enabled the capture of subtle relationships between linguistic content and delivery that were previously out of reach. These developments make this an opportune moment to revisit MAM in light of contemporary MMDL research.

However, transferring multimodal techniques directly to argument mining remains non-trivial. Arguments possess logical and structural properties that distinguish them from other multimodal tasks, such as sentiment analysis or visual question answering. Moreover, the field currently lacks sufficient multimodal datasets, standardised benchmarks, and dedicated tools or pre-trained models, all of which are necessary to develop and evaluate multimodal argumentative systems systematically. This scarcity of data and infrastructure, combined with the inherent complexity of aligning acoustic and textual features, remains a significant obstacle to progress in MAM research.

3.1.1 Challenges

Data Scarcity and Quality Issues — The most fundamental challenge has been the scarcity of high-quality multimodal corpora specifically designed for AM tasks. While the NLP community has developed numerous large-scale text corpora for various AM tasks, creating multimodal argumentative datasets presents unique challenges. Processing and annotating audio data is significantly more costly and time-consuming than working with text alone. Audio annotation requires specialised expertise, sophisticated tools, and careful attention to temporal alignment

between spoken content and written transcripts. Furthermore, privacy and copyright concerns often limit the availability of audio recordings from political events, news broadcasts, and other rich sources of argumentative discourse. To illustrate the severity of this limitation, at the beginning of this research, only a handful of MAM corpora existed. UKDebates [177], which addresses the task of claim detection, and M-Arg [207], which focuses on argumentative relations between sentences, were among the only available resources. These corpora, while valuable, were characterised by significant limitations: they were small-sized datasets featuring only a handful of speakers debating on one or a few occasions over a limited time span. Such constraints severely limited the generalisability of models trained on these datasets. They made it challenging to study phenomena such as speaker adaptation, the temporal evolution of political discourse, and cross-event generalisation. The contrast with text-only resources was stark. For instance, USElecDeb60To16 (USED), curated by Haddadan et al. [103], represented a significant advance in terms of scale and temporal coverage, featuring a substantial number of US presidential candidates debating over several decades. However, this valuable resource contained only annotated transcripts, with no links to the corresponding audio sources, thereby representing a missed opportunity for multimodal research.

Lack of Standardised Infrastructure, Tools, and Evaluation Frameworks — The second major challenge was the absence of standardised infrastructure and evaluation frameworks for MAM research. This problem mirrors similar issues in other rapidly evolving AI domains [165, 111], but the specific requirements of AM make it particularly acute.

MAM researchers struggled to access, process, and evaluate models and datasets due to fragmented resources scattered across various websites, repositories, and research group servers, each using distinct, often incompatible methods for data loading, preprocessing, and reconstruction. This created several problems: researchers spent disproportionate time on technical infrastructure rather than core research questions; fair comparisons between approaches became difficult; reproducing results was often challenging or impossible; and new researchers faced substantial barriers to entry.

The heterogeneous nature of multimodal data further complicated the situation. Unlike text-only AM, where standardised formats and preprocessing pipelines had emerged, multimodal systems required careful consideration of tem-

poral alignment, audio quality, feature extraction, and modality-specific preprocessing. The lack of shared tools meant each research group developed custom solutions, leading to inconsistencies that made cross-study comparisons unreliable.

Beyond infrastructure, the field lacked comprehensive evaluation frameworks and community coordination mechanisms. Unlike more established NLP areas that benefit from regular shared tasks, workshops, and standardised evaluation protocols, MAM research remained fragmented, with individual groups working in isolation. This made it difficult to assess the state-of-the-art, led to duplicated efforts, created barriers for identifying effective approaches, and limited the field's ability to attract new researchers and maintain momentum.

Limited Task Coverage and Unexplored Applications — A third significant challenge was the limited scope of tasks that had been explored in MAM contexts. While the field had made progress on core AM tasks such as argument detection, argument component classification, and relation classification [177, 348, 255], many important and potentially impactful applications remained unexplored. Particularly notable was the absence of work on Argumentative Fallacy Classification (AFC) [94] in multimodal settings. This gap was especially surprising given the strong theoretical motivation for applying multimodal approaches to fallacy detection. Research in rhetoric and sociolinguistics has established clear connections between acoustic features and fallacious reasoning patterns. For example, Kišiček [144] demonstrated how stereotypes about accents could be exploited in ad hominem fallacies, how staccato speech rhythms could be associated with inappropriate appeals to authority, and how prosodic elements emphasising anger could enhance the persuasive impact of appeals to emotion. These observations suggested that AFC should naturally be formulated as a multimodal task, yet no empirical work had been conducted to validate this hypothesis or develop appropriate computational methods. This represented not only a missed research opportunity but also a practical limitation, given the increasing importance of automated fact-checking and misinformation detection in political discourse.

3.1.2 Our Contribution

This chapter presents four interconnected contributions that address the aforementioned challenges and advance the state-of-the-art in multimodal argument mining:

- We develop and release MM-USED, an extended version of the USED corpus

that aligns textual input with corresponding audio segments, creating the largest multimodal dataset for argument mining in political debates. MM-USED supports multiple core tasks, including sentence detection, component and claim classification, and argumentative relation identification. It spans the widest temporal range and offers the greatest speaker diversity in its domain. Alongside the dataset, we conduct an extensive experimental study comparing traditional feature-based audio encodings with modern MMDL approaches, establishing the first comprehensive benchmarks for multimodal argument mining.

- We extend the investigation to AFC by introducing MM-USED-fallacy, the first corpus for multimodal AFC. Building on the MM-USED infrastructure, we demonstrate that integrating the audio modality leads to superior classification performance, confirming that framing fallacy classification as a multimodal task is essential to capture paralinguistic aspects of fallacious arguments.

- We develop and release MAMKit, a comprehensive PyTorch toolkit tailored for multimodal argument mining. The toolkit includes multiple datasets and models, providing researchers with a standardised platform for experimentation and evaluation. MAMKit provides a simple interface for loading, reconstructing, and processing existing datasets, facilitating seamless integration and comparison across tasks and datasets.

- We organise the first shared task on multimodal argumentative fallacy detection and classification in political debates. This initiative provides a platform for developing and evaluating systems that detect and classify argumentative fallacies across different modalities, fostering collaboration and innovation in the field while establishing standardised evaluation benchmarks.

3.1.3 Chapter Structure

This chapter is organised as follows, reflecting the chronological order of publication and the progressive development of our research program.

We begin with a review of the related work in Section 3.2, providing context for the challenges and existing approaches in multimodal argument mining and establishing the theoretical and technical foundations that motivate our research contributions.

In Section 3.3, we present our foundational exploration of multimodal argument mining, where we introduce the MM-USED corpus and conduct a comprehensive study of multimodal architectures across multiple core AM tasks, including sentence classification, component classification, claim detection, and argumentative relation classification. This section details the methodology used for creating the largest multimodal dataset for argument mining, the process of aligning audio inputs with textual annotations, and extensive architectural evaluations that establish foundational benchmarks for the field.

The chapter continues in Section 3.4 with our extension to the specific task of fallacy classification, where we introduce the MM-USED-fallacy corpus as the first multimodal resource for argumentative fallacy detection and classification.

In Section 3.5, we present our effort to standardise and democratize access to multimodal argument mining resources through the development of MAMKit, a comprehensive toolkit. This section details the design and implementation of the toolkit, which integrates multiple datasets and models while providing researchers with a standardised platform for experimentation, evaluation, and fair comparison across tasks and datasets. Additionally, we explore and integrate advanced audio encoding strategies and novel fusion architectures that extend beyond existing approaches in the literature, contributing new methodological insights to the field alongside the standardisation efforts.

Section 3.6 describes our organisation of the first shared task in multimodal argumentative fallacy analysis. We present the task design, evaluation framework, participating systems, and results that establish new benchmarks for the field, foster community collaboration, and provide standardised evaluation protocols for future research.

Finally, we present the conclusions in Section 3.7, summarising the main contributions of our comprehensive research program and outlining future directions for multimodal argument mining, including potential applications and methodological advances that can build upon the foundations established in this work.

Each section builds upon the previous contributions while advancing specific aspects of multimodal argument mining, collectively forming a comprehensive research program that has significantly contributed to establishing and advancing this emerging field.

3.2 Related Work

The foundations of MAM rest on several interconnected research areas that have evolved over the past decades. This section provides a comprehensive overview of the theoretical foundations, technical developments, and existing resources that contextualise our contributions.

3.2.1 Foundations

Argument mining has emerged as a significant area within computational linguistics and AI, focusing on the automatic identification, extraction, and analysis of argumentative structures from natural language text.

The field encompasses various tasks, including argumentative component detection and classification, relation identification, stance classification, and the analysis of argumentative discourse across different domains [163].

Traditional argument mining research has primarily focused on textual analysis, developing sophisticated methods for understanding how arguments are constructed, linked, and presented in written discourse. A crucial insight that motivated the development of MAM was the recognition of a strong connection between the argumentation process and the emotions of those involved. This observation led to the hypothesis that para-linguistic elements encoded in audio data are significant indicators that might aid in identifying arguments made in debates. Research on the interplay between arguments and emotions in speech began with Benlamine et al. [24], establishing the theoretical groundwork for incorporating acoustic features into argumentative analysis.

Subsequently, further studies focused on multimodality in argumentation, demonstrating the correlation between paralinguistic features and argumentative discourse across domains, including advertisements, news coverage, and legal analytics [143, 98, 144]. These findings revealed that argumentative discourse extends beyond textual content to encompass prosodic, acoustic, and other paralinguistic elements that can provide crucial information for understanding argumentative intent and structure.

Within this broader context, the study of fallacies represents a significant application area. The study of fallacies is deeply rooted in argumentative theory, dating back to Aristotle. The utility of recognising and studying fallacies emerged prominently in the 1970s-80s [104], driven by the understanding that fallacy de-

tection is directly related to human reasoning, where communication can often degenerate into conflicts, disagreements, and debates due to logical fallacies in discourse [126]. Detecting and classifying fallacies in discourse has proven valuable across several applications, including analysing human behaviour in dialogical settings, preventing the spread of misinformation in fact-checking systems, and evaluating generative models' reasoning.

However, fallacies are rarely expressed solely through language. They often rely on paralinguistic and prosodic cues, such as tone, emphasis, or rhythm, that reinforce or distort the logical content of speech. For instance, an appeal to emotion or an ad hominem attack may be conveyed as much by how something is said as by what is said. This makes fallacy detection a naturally multimodal task, where analysing both text and audio is essential to capture the interplay between rhetorical form, emotional delivery, and argumentative structure.

3.2.2 Prior Multimodal Argument Mining Research

Work in MAM started relatively recently, inspired by studies on the connections between arguments and emotions [24].

The first significant contribution came from Lippi et al. [177], who developed a classifier for claim detection from speech in the domain of political debates using a televised debate from the 2015 UK political elections. They built a first-of-a-kind political debate corpus by annotating arguments from three prime ministerial candidates and demonstrated that audio features improve claim detection when used as input to a Support Vector Machine (SVM) classifier alongside textual transcripts.

Subsequent research expanded these initial efforts. Mestre et al. [207] built the M-Arg corpus, consisting of 4,104 labelled sentence pairs selected from debates of the 2020 US political elections, and experimented with different multimodal input models. Outside of political speeches, Mirkin et al. [211, 212] developed a corpus that couples transcripts and audio from several debates, though these are non-political debates carried out by paid actors on controversial topics. The interest in political debates motivated further research and the development of resources. More recently, Mestre et al. [206] extended their earlier work by augmenting MAM approaches and exploring argumentative relation identification tasks. At the time this research began, these small-scale corpora represented the extent of available multimodal resources for argument mining.

Significant limitations characterised the UKDebates corpus [177] and M-Arg [207]: they featured only a handful of speakers debating on a few occasions over limited time spans, severely constraining the generalizability of models and making it challenging to study phenomena such as speaker adaptation or the temporal evolution of political discourse.

Similar efforts have been conducted in the context of fake news detection, where Ivanov et al.[121] observed superior classification performance in several tasks, such as the detection of check-worthy claims, when following a multimodal problem formulation.

Research has also established clear connections between acoustic features and fallacious reasoning patterns, providing strong theoretical motivation for applying multimodal approaches to fallacy detection. Studies have shown how stereotypes about accents can be exploited in ad hominem fallacies, how staccato speech rhythms can be associated with inappropriate appeals to authority, and how prosodic elements emphasising anger can enhance the persuasive impact of appeals to emotion [144]. However, despite these promising initial findings, MAM research remained severely limited by several factors: the difficulty and heterogeneity of AM tasks, the scarcity of large-scale multimodal AM data, and the inherent challenges of MMDL.

3.2.3 Multimodal Deep Learning for Multimodal Argument Mining

Recent developments in MMDL have opened promising directions for advancing MAM. These include transfer learning frameworks that mitigate data scarcity [224] and attention-based architectures that model interactions within and across modalities [171, 337, 99].

The use of pre-trained models and transfer learning is particularly crucial in MAM, where datasets are typically small, often comprising hundreds or a few thousand samples rather than the millions required to train deep neural networks from scratch. By leveraging models trained on large, general-purpose corpora, transfer learning provides rich, semantically informed representations that can be fine-tuned or used as fixed feature extractors for specific argumentative tasks. This strategy effectively transfers broad multimodal knowledge to the resource-constrained MAM setting.

A representative example is AudiBERT [332], which integrates pre-trained text and audio encoders through a dual self-attention mechanism, illustrating the

benefits of transfer learning for multimodal integration. More generally, modern MMDL architectures combine unimodal encoders through diverse fusion strategies, allowing each modality to contribute specialised information before integration.

However, existing MAM systems have yet to fully exploit the capabilities of these modern transfer learning frameworks. Most rely on conventional feature concatenation rather than on the representational depth offered by pre-trained encoders. Recent advances in MMDL now enable more sophisticated approaches, including powerful text encoders such as RoBERTa [179] and SBERT [274], and audio encoders such as CLAP [363], Wav2Vec 2.0 [11], HuBERT [117], and WavLM [46]. These models provide a foundation for transfer learning strategies that can better capture the intricate relationships between textual and acoustic cues in argumentative discourse, while remaining effective in the data-limited conditions typical of MAM.

3.2.4 Standardization and Tools

In recent years, there has been a growing emphasis on streamlining training, evaluation, and benchmarking processes across diverse AI domains. Several resources have become available to address specific tasks and applications. For benchmarking and multimodal research, LAVIS [165], MMF [303], X-modaler [169], and UniLM [208] provide user-friendly interfaces for accessing datasets and for training/evaluating language-vision models. Furthermore, several specialised tools have been proposed for various multimodalities, including TorchMultimodal [257] for accessing state-of-the-art multimodal models, ViLMedic [67] for vision and language in medical AI, pyannote.metrics [31] and pyannote.audio [32] for speaker diarization, and Muskits [299] for end-to-end music processing. Specialized NLP libraries and tools focus on specific tasks, including LogiTorch [111] for logical reasoning in natural language, TextBox 2.0 [321] for text generation using pre-trained language models, mahaNLP [189] for Marathi NLP, DeepPavlov [34] for dialogue systems, TextAttack [219] and OpenAttack [379] for adversarial attacks in NLP, LambdaKG [365] for knowledge graph embeddings, nerblackbox [315] for named entity recognition, NewsRecLib [119] for news recommendation, and NeuralQA [73] for question answering. Despite this rich ecosystem of tools and frameworks in related areas, to the best of our knowledge, no such resources exist to support argument mining or specifically MAM research. This absence of stan-

dardised tools and evaluation frameworks represents a significant gap that has hindered progress in the field, as researchers must develop custom solutions for data processing, model implementation, and evaluation, making fair comparisons and reproducible research challenging.

3.2.5 Research Gaps

Several significant gaps emerge from this review of related work. First, while MAM has shown promise in core tasks like argument detection and component classification, many important applications remain unexplored, and more recent tasks like fallacy classification and detection are largely underexplored in multimodal settings. Second, the field lacks large-scale, comprehensive datasets that can support robust model development and evaluation across multiple tasks. Third, despite advances in multimodal deep learning, MAM research has not fully leveraged state-of-the-art architectural innovations and encoding strategies. Finally, the absence of standardised tools and evaluation frameworks has fragmented the research community and limited the field's growth potential. These gaps motivate the work presented in this chapter, which addresses the challenges of data scarcity, methodological limitations, and the lack of standards and tools for collaborative research in MAM.

3.3 Multimodal Argument Mining: A Case Study in Political Debates

This section presents our foundational work in MAM [197], conducted in collaboration with Federico Ruggeri, Andrea Galassi, and Paolo Torroni.

Key Contribution: We introduce MM-USED, the largest multimodal corpus for argument mining at the time of its release, and conduct a comprehensive study of multimodal architectures across core argument mining tasks.[a] This work establishes important benchmarks for the field and provides insights into audio encoding strategies and multimodal fusion challenges that inform the subsequent contributions presented in this chapter.

[a]Code and resources available at `https://github.com/nlp-unibo/multimodal-am`

We consider four distinct classification tasks to provide comprehensive coverage of core argument mining capabilities:

- **Argumentative Sentence Detection (ASD)**: an input sentence x is classified as containing an argument (*arg*), or not containing an argument (*not-arg*);
- **Argumentative Component Classification (ACC)**: an argumentative sentence x is classified as containing a *claim* or a *premise*;
- **Claim Detection (CD)**: a sentence x is classified as containing a *claim* or not containing a claim (*not-claim*);
- **Argumentative Relation Classification (ARC)**: a pair of sentences x_i and x_j is classified as yielding an argumentative relation $x_i \rightarrow x_j$ of *support*, *attack*, or *neither* (if no argumentative relation exists).

Each input is characterized by two modalities: the textual input x_t and the audio input x_a. To assess the impact of each modality, we consider three distinct input configurations: *Text-Only (TO)*, *Audio-Only (AO)*, and *Text-Audio (TA)*, where both modalities are given as input.

3.3.1 Datasets and Corpus Development

We experiment on three different debate corpora, designed to address four separate but strictly correlated AM tasks. Table 3.1 summarizes the corpora's key figures.

Corpus	Sentences	Debates	Speakers	Years	Class Distribution	Task(s)
UKDebate [177]	386	1	3	2015	152 claim, 234 not-claim	CD
M-Arg [207]	4,104 pairs	5	4	2020	120 attack, 384 support, 3600 neither	ARC
M-Arg ($\gamma \geq 0.85$) [207]	2,443 pairs	5	4	2020	29 attack, 132 support, 2282 neither	ARC
MM-USED (Ours)	26,781	39	26	1960-2016	10,882 claim, 9,683 premise, 6,226 not-arg	ASD, ACC

Table 3.1: *Corpora for MAM. For M-Arg, we also consider the corpus version in which samples have high annotation confidence, γ.*

UKDebates Corpus — UKDebates, by Lippi et al. [177], was the first corpus released for MAM. Its context is the UK Prime Ministerial elections of 2015. It is based on the two-hour debate aired by Sky News on April 2, 2015, and it comprises the audio sequences of 3 candidates: David Cameron, Nick Clegg, and Ed Miliband. UKDebates contains 386 audio samples (122 for David Cameron, 104 for Nick Clegg, 160 for Ed Miliband) of varying length, accompanied by a human-built transcript. Two domain experts annotated the collected transcripts for the task of claim detection by labelling each sentence as containing or not containing a claim. Regarding Inter-Annotator Agreement (IAA), the authors report $\kappa = 0.53$, "fair to good" agreement. Because audio features are markedly speaker-dependent, Lippi et al. address the CD task for each politician candidate in turn. The authors report a F1-score of ∼59-62%.

M-Arg Corpus — M-Arg, by Mestre et al. [207], is built around the 2020 US Presidential debates. The debates feature 5 speakers (4 candidates and a moderator) and cover 18 topics. A carefully designed crowd-sourcing exercise resulted in 4,104 labelled sentence pairs for the task of argumentative relation detection. In particular, each sentence pair was labelled as *support*, *attack*, or *neither*. To account for crowd-workers' annotation quality, each label is enriched with an *annotator agreement confidence* γ. A smaller but higher-quality subset of M-Arg is thus obtained by only selecting the links with confidence $\gamma \geq 0.85$. The price of this reduction in annotations' noise is a smaller dataset, which results in harder training and an IAA of $\alpha = 0.43$. Mestre et al. report a macro F1-score of 22.5% and 11.0% for the ARC task regarding the whole corpus and the ($\gamma \geq 0.85$) subset, respectively. The macro F1-score regards the *support* and *attack* labels only.

USED Corpus — USED, by Haddadan et al. [103], is the largest collection of annotated textual documents for AM in the political debates domain. It contains transcripts of presidential candidates' debates aired from 1960 to 2016. Annotations are at the sentence level. Each sentence is labelled as a claim, a premise, or neither of them. The authors used this corpus to address the ASD and ACC tasks. Regarding IAA, they report a $\kappa = 0.57$ (moderate agreement) for ASD and a $\kappa = 0.40$ (fair agreement) for ACC. As for classification performance, Haddadan et al. [103] report a macro F1-score of 73.0% and 76.95% for the ASD and ACC, respectively.

MM-USED Corpus — We built MM-USED by augmenting USED with the audio modality. We remark that we did not add any additional labels, nor did we modify existing ones. We obtained the audio files of the debates from the PBS NewsHour YouTube channel.[1] Before aligning transcripts with corresponding audio files, we conducted a preliminary preprocessing step. First, we manually trimmed audio files to remove content not included in the paired transcripts, such as the moderators' opening and closing remarks. In some cases, audio files can contain cuts spanning from a few seconds to several minutes. We removed the sentences from the transcripts that matched these cuts. Second, we removed transcripts that did not match their paired audio files or were incomplete. Third, we removed metadata, such as the speaker's information, from each transcript to avoid spurious alignments. Fourth, we tokenised transcripts; thus, the resulting transcripts contain one sentence per line. See Appendix A.1 for additional details.

After pre-processing, we split each audio file into 20-minute chunks to improve the alignment quality. We manually extracted the text from the transcripts corresponding to the created audio files. We used Aeneas[2] to automatically retrieve the start and end timestamps of each utterance. Lastly, we post-processed our corpus by removing (i) sentences misaligned with their audio sample, (ii) sentences not matching any of the aligned utterances (e.g., transcription tags like "applause"), (iii) non-argumentative duplicated sentences, such as *Thank You* or *Ok*. Finally, we verified the quality of the alignments by spot checks. In particular, we checked several parts of each debate and found no major misalignments.

As a result of the mentioned pre- and post-processing phases, we removed about 2,000 samples from the original USED corpus. The resulting MM-USED

[1] `https://www.youtube.com/channel/UC6ZFN9Tx6xh-skXCuRHCDpQ`
[2] `https://github.com/readbeyond/aeneas/`

corpus contains 26,791 annotated textual sentences and their corresponding audio samples.

Our corpus differs from previous MAM corpora in terms of size, variety and annotation quality. First off, it is the largest MAM corpus to date, by a significant margin. Second, it offers a wider range of speakers over a much longer time span (1976-2016), possibly paving the way to new research perspectives, such as the analysis of the evolution of political communication, argumentation and language over time. The greater number of different speakers could also facilitate the creation of more robust classification models. Finally, the corpus includes expert annotations rather than crowd-sourced ones.

3.3.2 Methodology and Experimental Design

Architectural Framework

We define three models based on the high-level schema illustrated in Figure 3.1. In all our models, each modality is processed separately by either a *text module* or an *audio module*. Each module is part of a classification model defined for a particular input modality. Different input configurations use other modules. The TO and AO input configurations consider only the text or audio module, respectively. In the TA multimodal setting, instead, the outputs of the two modules are concatenated and passed through a final *classification module*. The classification module receives an encoded representation of one or more modalities, depending on the input configuration, and produces a classification label. For each model, we experiment with two different audio signal encoding methods: a set of widely-adopted spectral features [275] and the Wav2vec embeddings [289]. Such encoding methods represent a preliminary pre-processing step of the audio signal, which is then passed as input to the audio module.

The models are defined as follows.

- **SVM** follows Lippi et al. [177]. The text module encodes input textual sentences as TF-IDF vectors. The audio module is an identity function, that is, the encoded audio signal remains unaltered. The classification module is an SVM classifier.
- **M-ArgNet** reflects the neural architecture presented in [207]. The text module is defined by a pre-trained BERT [70] model. The audio module is a stack of Convolutional Neural Network (CNN) layers with a Bidirectional

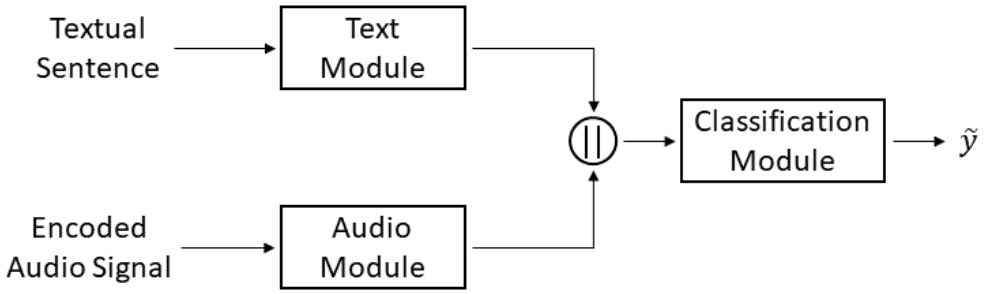


Figure 3.1: *The proposed schema for MAM.*

Long Short-Term Memory (BiLSTM) layer on top. The classification module is a Multilayer Perceptron (MLP).

- **BiLSTM** is a third architecture where the text module comprises a pre-trained GloVe [250] embedding layer to encode input textual sentences and a stack of BiLSTM layers. The audio module is defined by another stack of BiLSTM layers. The classification module is an MLP.

M-ArgNet and BiLSTM, when used with Wav2vec embeddings, resemble AudiBERT [332] since they are all based on text and audio embedding taken from pre-trained models.

In addition to the above models, we also consider a weighted random baseline classifier, i.e. **Random**, which acts as a lower bound for each task of interest.

3.3.2.1 Audio Representation Strategies

In this section, we provide additional details regarding the described audio signal encoding methods. The first method, denoted as *feature-based* encoding, is a set of widely-adopted spectral features [275], such as the MFCCs, spectral centroids, spectral bandwidth, spectral roll-off, spectral contrast and a 12-bit chroma vector. The result of this extraction process is a two-dimensional feature matrix of shape (no. frame, no. features). Following Mestre et al. [207], we consider 25 MFCCs and 20 other spectral features, for a total of 45 features.[3] Regarding the number of frames, their amount is proportional to the duration of the audio signal. In our experimental setup, it is in the order of hundreds. To reduce the number of frames, we adopt average pooling. This applies a moving average with a parametric window size on the frame dimension. We experiment with different window sizes

[3]We used the `librosa` library with default parameters.

to reduce the computational demand and the number of parameters of our models, without degrading the informative content of the audio signal.

The second method, denoted as *embedding-based* encoding, uses the end-to-end audio encoding neural architecture Wav2vec [289].[4]

In particular, we directly extract the pooled embedding vector given by the model. The final size of the representation is a 768-dimensional embedding vector according to the chosen Wav2vec model.

Experimental Protocol

We define a reproducible and robust experimental setup to evaluate the contribution of each modality to AM tasks, and to assess the impact of different input representations and classifiers. The limited amount of data, especially in UKDebates and M-Arg, caused our setup to differ in several ways from previous studies. Hence our results are not directly comparable with those published in the relevant literature. Nonetheless, our setting includes all the classifiers used in such studies, in addition to more recent representation techniques.

Regarding UKDebates, Lippi et al [177] address the CD task by experimenting on each politician individually using a 10-fold cross-validation routine. In contrast, we evaluate our models on all speaker sentences via a repeated 5-fold cross-validation routine to curb the high variance typically observed when neural models are trained with limited data [21].

Similarly, for M-Arg, we use repeated 5-fold cross-validation rather than the train-validation split approach proposed by Mestre et al. [207].

For both datasets, we set the number of repetitions to 3.

For MM-USED we follow the same experimental setup as in [103]. Despite the different number of samples, we keep the same train, validation, and test splits proposed for the original corpus. We define a repeated training and evaluation routine for model benchmark, setting the number of repetitions to 3. See Appendix A.2 for additional details on our experimental setting, number of samples and data splitting.

Optimization and Training

We train our neural models using cross-entropy as the optimization objective and Adam [142] as an optimizer. Additionally, we regularize neural models by

[4]We use the *facebook/wav2vec2-base-960h* model version.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	<u>66.24</u>	48.62	49.13	46.20	**66.71**
M-ArgNet	<u>67.20</u>	47.20	65.94	50.12	**68.68**
BiLSTM	<u>66.81</u>	45.40	65.29	50.20	**66.84**
Random			40.90		

Table 3.2: *Average binary F1-score regarding the claim class on the test set of UKDebates. For each row, we report the best results in bold, second best results are underlined instead.*

applying early stopping on the validation loss with patience set to 10 epochs and using dropout [313].

All models undergo a preliminary hyper-parameters calibration phase. In particular, for each input configuration (i.e., TO, AO, and TA) we calibrate the models to assess the contribution of individual modalities. Additional details about model calibration are reported in Appendix A.3.

3.3.3 Results and Analysis

We report the classification results on each dataset. Additionally, we perform an ablation study regarding the models trained in the TA configuration to evaluate the contribution of each input modality.

UKDebates Results — Table 3.2 reports classification results for the CD task on the UKDebates corpus. In particular, we compute the average binary F1-score on the test set for each input configuration and audio encoding method. We provide the F1-score as a customary performance indicator in unbalanced classification situations. We observe that the best-performing input configuration for each model is the TA with embedding-based audio encoding. However, the gap with respect to the TO input configuration is marginal, suggesting that the audio modality is not efficiently handled by the employed models or is not sufficiently informative. Regarding audio encoding, we observe that the embedding-based method leads to better performance than the feature-based approach. This is evident for the SVM classifier, where the TA setting with embedding-based audio encoding leads to an improvement of more than 17 F1-score percentage points compared to its feature-based counterpart.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	<u>14.70</u>	11.96	12.33	14.09	**16.73**
M-ArgNet	16.24	<u>18.45</u>	18.27	8.88	**19.02**
BiLSTM	<u>16.78</u>	9.18	15.89	9.84	**20.21**
Random			2.79		

Table 3.3: *Average macro F1-score concerning the attack and support classes on the test set of M-Arg ($\gamma \geq 0.85$). For each row, we report the best results in bold, second best results are underlined instead.*

M-Arg Results — Table 3.3 reports the average macro F1-score regarding the *attack* and *support* classes on the test of the M-Arg corpus for the ARC task. We focus on the M-Arg corpus version with annotation confidence $\gamma \geq 0.85$ to consider high-quality examples only. We observe that the TA input configuration with the embedding-based audio representation is the best performing one for all the considered classifiers. In particular, such a configuration outperforms the TO input configuration by 2.03, 2.78 and 3.43 F1-score percentage points for SVM, M-ArgNet, and BiLSTM classifiers, respectively. In contrast, the TA input configuration with feature-based audio representation yields mixed results. More precisely, only the M-ArgNet model outperforms its TO counterpart. This is in agreement with the results obtained in CD on UKDebates. The feature-based AO input configuration is remarkably on par with its TA counterpart.

MM-USED Results — Table 3.4 reports classification performance concerning the ASD and ACC tasks evaluated on the test set of the MM-USED corpus. In general, the embedding-based audio encoding appears to perform better than the feature-based one. This agrees with the behaviour observed in the previous experiments. However, we observe that the TA configuration does not always perform better than TO. We hypothesize that the characteristics of this corpus, with multiple speakers spanning several decades, bring in additional challenges that these architectures are not addressing effectively. Concerning ASD, we observe that the TA input configuration is the best performing one for the BiLSTM and the M-ArgNet models. In contrast, the TO input configuration leads to superior performance for the SVM model. Overall, there is no significant performance gap between the TA and TO input configurations. However, the AO input configurations with both audio signal representations are not far behind their TO and TA counterparts. All this suggests that the encoded audio signal is informative

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
ASD					
SVM	**67.18**	49.37	49.02	61.20	<u>65.38</u>
M-ArgNet	<u>65.64</u>	52.71	60.89	61.04	**68.38**
BiLSTM	67.19	58.89	**68.57**	60.40	<u>68.23</u>
Random			50.54		
ACC					
SVM	**65.85**	50.19	51.66	58.44	<u>64.75</u>
M-ArgNet	**67.40**	50.05	60.09	65.33	<u>67.38</u>
BiLSTM	<u>65.99</u>	49.58	**66.25**	58.86	65.80
Random			50.51		

Table 3.4: *Average macro F1-score on the test set of MM-USED. For each row, we report the best results in bold, second best results are underlined.*

to address the task, but the fusion of both modalities is non-trivial depending on the given audio representation. We observe similar behaviours concerning the ACC task. In particular, the TA input configurations do not lead to consistent performance benefits for the employed models. Nonetheless, the AO input configuration with embedding-based audio representation significantly outperforms its feature-based counterpart. These observations confirm a known fact, that merging multiple input modalities is still a major challenge in current multimodal models.

Ablation Study — To assess the contribution of each individual input modality, we carried out an ablation study on models trained with the TA input configuration, by alternatively masking either input modalities. To this end, we zeroed out the output embedding vector of the input module corresponding to the modality to be masked.

Table 3.5 reports the results of the ablation study regarding the UKDebates corpus.

We observe that the BiLSTM model with feature-based audio representation reaches the same performance in both the TA and TO (i.e., w/o Audio) configurations. This result suggests that the audio modality does not provide informative content in addition to text for the task. From a reversed perspective, the SVM classifier with feature-based audio representation focuses solely on the audio modality. We interpret this as an effect of the difficulty of combining text and audio modalities for the SVM classifier. We observe similar behaviours when

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	65.29	49.13	65.94
w/o Text	21.00	49.13	46.04
w/o Audio	65.29	0.00	57.02
Embedding-based			
TA	66.84	66.71	68.68
w/o Text	3.81	16.40	11.27
w/o Audio	66.78	0.00	68.48

Table 3.5: *Ablation test regarding TA model configuration on the UKDebates test set.*

considering the embedding-based audio representation as well. In contrast, the M-ArgNet model behaves in line with our initial expectations regarding the ablation study. In particular, the model achieves superior performance compared to the random baseline when one of the input modalities is removed, while being inferior to the default TA case. The only exception concerns the embedding-based audio representation setting. In this setting, the text modality significantly contributes to the task compared to the audio modality.

Likewise, with the M-Arg corpus (see Table 3.6), we observe odd results similar to those observed with UKDebates.

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	15.89	12.33	18.27
w/o Text	0.0	12.33	9.21
w/o Audio	10.00	0.00	3.78
Embedding-based			
TA	20.21	16.73	19.02
w/o Text	0.0	9.30	1.16
w/o Audio	12.80	0.00	18.24

Table 3.6: *Ablation test regarding TA model configuration on the M-ARG ($\gamma \geq 0.85$) test set.*

In particular, the BiLSTM and SVM models show symmetrical effects concerning performance metrics when one of the input modalities is removed. Independently of the audio representation method, the BiLSTM model heavily relies on text information to perform the task. In contrast, the SVM model fails to ad-

dress the task when audio is removed. This evidence suggest that the way input is encoded also plays an important role in a multimodal model concerning the impact of each modality.

Furthermore, we observe similar issues in the MM-USED corpus when addressing the ASD and ACC tasks. Table 3.7 reports the results of the ablation study concerning the ASD task. Again, we observe that the BiLSTM and SVM models have symmetric behaviours. Additionally, the BiLSTM reaches superior classification performance when removing the audio modality in both audio representation settings. Despite a small improvement, this surprising result suggests that the audio modality might be noisy and, thus, detrimental to the task. This observation is further supported by the low performance achieved when removing the text modality. We observe this phenomenon also in the ACC task as reported in Table 3.8. In particular, the M-ArgNet with embedding-based audio representation has superior performance when removing the audio modality compared to the default TA input configuration.

Input	**BiLSTM**	**SVM**	**M-ArgNet**
Feature-based			
TA	68.57	49.02	60.89
w/o Text	17.26	49.02	23.11
w/o Audio	69.40	17.26	48.04
Embedding-based			
TA	68.23	65.38	68.38
w/o Text	17.26	61.11	44.35
w/o Audio	68.44	17.26	33.95

Table 3.7: *Ablation for ASD task*

Input	**BiLSTM**	**SVM**	**M-ArgNet**
Feature-based			
TA	66.25	51.66	60.09
w/o Text	32.71	51.66	44.25
w/o Audio	66.24	32.71	55.25
Embedding-based			
TA	65.80	64.75	67.38
w/o Text	32.71	48.86	33.95
w/o Audio	65.80	33.57	67.57

Table 3.8: *Ablation for ACC task*

3.3.4 Discussion and Findings

The results presented so far highlight several key trends in MAM. Across all datasets, embedding-based audio encodings consistently outperform traditional feature-based representations, in line with recent advances in MMDL [289], confirming that leveraging pretrained audio embeddings is a promising direction for improving robustness and generalisability. The text-audio (TA) configuration, which integrates linguistic and acoustic information, generally achieves performance comparable to or higher than text-only (TO) models, reinforcing the importance of prosodic and paralinguistic cues in argumentative analysis. Interestingly, audio-only (AO) configurations demonstrate that speech signals alone convey

meaningful argumentative information, offering a valuable solution when textual transcriptions are unavailable or unreliable. However, defining effective strategies for input encoding and multimodal fusion remains a central challenge. Our results indicate that current architectures only partially capture the complex interactions between modalities, suggesting that attention-based and intermediate-fusion approaches from recent multimodal learning research could yield further improvements.

To enable systematic investigation of these challenges, we released MM-USED, the largest dataset to date for MAM, comprising U.S. presidential debates and speeches. We conducted a comprehensive empirical study across three heterogeneous corpora differing in size, topic, annotation scheme, and speaker variety, establishing benchmarks that pave the way for future research in low-resource and cross-domain settings.

3.4 Multimodal Fallacy Classification in Political Debates

This section presents our extension to argumentative fallacy classification in multimodal settings [198], conducted in collaboration with Federico Ruggeri and Paolo Torroni.

Key Contribution: Building on the foundational work presented in Section 3.3, we introduce MM-USED-fallacy, the first corpus specifically designed for multimodal argumentative fallacy detection and classification.[a] This work demonstrates that framing fallacy classification as a multimodal task is essential to capture paralinguistic aspects of fallacious arguments, confirming theoretical motivations from rhetoric and sociolinguistics research. The findings provide crucial insights into how acoustic features complement textual analysis for detecting specific types of logical fallacies in political discourse.

[a]Code and resources available at `https://github.com/nlp-unibo/multimodal-am-fallacy`

3.4.1 Domain Description

The term *fallacy* indicates a deceptive, misleading, or generally invalid argument [104, 350]. While the study of fallacies dates back to Aristotle, computational detection and classification of fallacies in natural language represents a relatively recent but increasingly important research area within argument mining. This is particularly crucial for analysing political discourse, where fallacious reasoning can significantly impact public opinion and democratic processes.

Following Walton's comprehensive scheme of argumentation fallacies [350], which provides a systematic framework for categorising different types of invalid reasoning patterns, we focus on six specific fallacy categories that are particularly prevalent in political discourse. Our selection of these six categories is motivated by their frequent occurrence in the corpus we extend in this work (which we present in the subsequent section), where they serve as macro categories that encompass the most common patterns of flawed reasoning encountered in political debates.

The six fallacy categories we address are:

Appeal to Authority refers to the use of an expert's opinion as evidence to back up an argument when the authority cited is not relevant to the domain of the

argument or when the appeal bypasses substantive reasoning.

Ad Hominem fallacy is characterised by an excessive attack on an arguer's character, circumstances, or actions rather than addressing the substance of their argument.

Appeal to Emotion involves the manipulation of an audience's emotions through loaded language, vivid imagery, or emotional appeals that substitute for logical reasoning.

False Cause regards the misinterpretation of correlation as causation, or the attribution of causality based on insufficient evidence of a causal relationship.

Slogans are brief and striking phrases used to evoke excitement or emotional response while avoiding substantive argumentation or evidence.

Slippery Slope is an argument that claims exaggerated or unrealistic outcomes will inevitably follow from a particular action, without adequate justification for the claimed causal chain.

These categories represent the most frequent patterns of fallacious reasoning in political contexts and provide a comprehensive framework for analysing the logical validity of political arguments.

3.4.2 Structural Foundation

The development of computational approaches requires appropriate datasets for training and evaluation. USED [103], containing annotated transcripts of US presidential debates from 1960 to 2016, provided an ideal foundation for fallacy detection research due to its comprehensive coverage of political argumentation over multiple decades.

Recognising this opportunity, Goffredo et al. [94] developed USElecDeb60To16-fallacy (USED-fallacy) as the first large-scale corpus specifically designed for automatic fallacy detection and classification in political debates. This groundbreaking work addressed a significant gap in the argument mining field, where previous research had focused primarily on identifying argumentative components and relations rather than analyzing the logical validity of arguments themselves.

Their approach involved a systematic annotation process where domain experts identified and categorized fallacious reasoning patterns across the extensive debate transcripts in USED. Rather than attempting to cover all possible fallacy types from Walton's comprehensive scheme, Goffredo et al. strategically focused on the six categories described above, selected based on their frequency in political

argumentation and their potential impact on public persuasion.

A critical innovation in Goffredo et al.'s methodology was their recognition that fallacies have various formats and styles, often spanning multiple sentences without aligning with sentence boundaries. For instance, *Appeal to Authority* and *False Cause* typically span over multiple sentences, whereas *Ad Hominem* attacks could be limited to just a few words. This observation led them to adopt a span-level annotation approach that allows for precise localization of fallacious reasoning within broader argumentative contexts, rather than forcing artificial sentence-level classifications.

The comprehensive annotation structure of USED-fallacy provides rich contextual information that goes beyond simple fallacy labels. Each annotation includes the dialogue context, the specific fallacy snippet, subcategory information, and relationships to argumentative components, enabling sophisticated analysis of how fallacies function within broader argumentative structures. Table 3.9 details the annotation structure employed in USED-fallacy.

Annotated Element	**Description**
Dialogue	a debate portion in which the fallacy is present
Snippet	the annotated fallacy in the dialogue
Fallacy	the label of the fallacy
Subcategory	the subcategory of the fallacy
Component Text	the component text in which the fallacy is found or the closest one
Component Label	the type of component
Relation Label	the relation type between the component and the fallacy

Table 3.9: *Annotations in USED-fallacy.*

Table 3.10 shows examples of annotated fallacies in USED-fallacy, illustrating the diverse manifestations of fallacious reasoning in political debate contexts and demonstrating how each of the six fallacy categories appears in actual political discourse.

The creation of USED-fallacy thus represents a significant milestone in argument mining research, providing the computational linguistics community with the first large-scale, expertly annotated resource for studying fallacious reasoning in naturalistic political discourse.

Snippet	Fallacy Category
the same kind of woolly thinking	Appeal to Emotion
As George Will said the other day, "Freedom on the march; not in Russia right now."	Appeal to Authority
Governor Carter apparently doesn't know the facts.	Ad Hominem
We won the Cold War because we invested and we went forward.	False Cause
And if we don't act today, the problem will be valued in the trillions.	Slippery Slope
We have to practice what we preach.	Slogan

Table 3.10: *Examples of annotated fallacies.*

3.4.3 Corpus Development

Our corpus, MM-USED-fallacy, integrates MM-USED and USED-fallacy to create the first multimodal resource for AFC. To build it, we leverage two resources: span-level fallacy snippet and argument component annotations from USED-fallacy, and sentence-to-audio alignment timestamps from MM-USED. Given that the two resources present a mismatch in granularity, we map span-level annotations from USED-fallacy to the sentence level and then align them with audio recordings. Alternatively, one could rely on text/audio alignments at the word level. However, that would require producing a new resource. Since our aim is to evaluate fallacy classification from a multimodal perspective, we decided to build as much as possible on what was available. We leave word-level text-to-audio alignment to future work.

Span-to-sentence mapping — We label a (sequence of) sentence(s) to a fallacy category or argument component type if they overlap with a span labelled as such.

Audio alignment — We use the retrieval tool developed in Section 3.3 to download the audio files corresponding to the debates annotated in USED-fallacy. In this process, we exclude three debates from the corpus for which no audio recording is found (see Appendix B.3 for more details). We then use the text-to-audio alignment timestamps to extract audio clips corresponding to the dialogues, the fallacy snippets, and the argumentative components. We perform the alignment via fuzzy string matching since we notice slight transcript mismatches between USED-fallacy and MM-USED due to different text pre-processing.

In this process, we also observe a few inconsistencies between USED-fallacy and MM-USED, amounting to duplicate sentences, punctuation differences, and

text segments missing for USED-fallacy. For this reason, we perform additional text processing steps and remove misaligned samples (see Appendix B.4 for more details). Our final corpus, MM-USED-fallacy, comprises 1891 text-audio pairs on 28 debates.

3.4.4 Corpus Statistics

Table 3.11 compares MM-USED-fallacy with the original USED-fallacy corpus, showing the distribution of samples across the six fallacy categories. The reduction in sample size reflects the challenges of aligning textual annotations with audio recordings and the exclusion of debates for which audio was not available.

Fallacy	**USED-fallacy**	**MM-USED-fallacy**
Appeal to Emotion	1,427	1,102
Appeal to Authority	416	326
Ad Hominem	279	157
False Cause	179	154
Slippery Slope	118	102
Slogans	78	50
Total	2,497	1,891

Table 3.11: *Number of samples in USED-fallacy and MM-USED-fallacy.*

Table 3.12 reports sentence and audio distribution per dialogue, fallacy snippet, and argumentative component in MM-USED-fallacy. We observe that fallacy snippet length varies between one sentence and eight, indicating high annotation variability. The distribution of snippet length across fallacy categories is shown in Table 3.13.

	Min	Max	Mean	Std
Text				
Dialogue	16	449	120.92	84.44
Snippet	1	8	1.33	0.82
Component	1	6	1.02	0.23
Audio (in seconds)				
Dialogue	91.28	4271.40	608.68	582.98
Snippet	0.32	74.32	10.40	7.46
Component	0.32	49.68	8.98	6.51

Table 3.12: *Text and audio length distribution (unit: sentences).*

		Length							
Fallacy	**Span**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Appeal to Emotion	330	482	168	51	27	30	-	5	9
Appeal to Authority	18	162	79	17	30	7	5	8	-
Ad Hominem	15	55	38	13	15	15	6	-	-
False Cause	6	58	26	17	30	-	-	8	9
Slippery Slope	5	52	24	16	-	5	-	-	-
Slogans	35	15	-	-	-	-	-	-	-
Total	409	824	335	114	102	57	11	21	18

Table 3.13: *Snippet sentence length distribution per category (unit: sentences).*

We observe that 409 (21.6%) fallacy snippets are at the span level. We remark that these snippets are mapped to sentence-level annotations. Among the 658 snippets spanning over two or more sentences, 578 ($\sim$ 88%) are aligned with an exact match. Notably, the fallacy categories of *Appeal To Emotion*, *Appeal to Authority*, and *False Cause* have the longest snippets. *Appeal To Emotion* is also the most frequent class. In contrast, the *Slogan* category has the smallest number of samples. Lastly, some fallacy categories are not present in all debates. We expect that the low representation of certain categories across the debates will have a negative impact on classification accuracy.

Additional insights from the data show that the span length varies between a minimum of 1 word and a maximum of 57 words. Furthermore, 24 sentences, representing 2.2% of the entire dataset, have multiple associated snippets. This small fraction suggests that the agreement statistics closely align with those reported by Goffredo et al. [94], with minor variations attributed to this subset. Considering the limited occurrence of sentences with multiple associated snippets relative to the total number of snippets in the dataset, we believe that this form of annotation mapping does not introduce significant drawbacks.

Span-to-sentence coverage analysis — While the annotation mapping itself introduces minimal label ambiguity (as discussed above), the change in granularity from span to sentence has a second, distinct consequence for the audio modality: when a short fallacious span is mapped to a longer reference sentence, the corresponding audio clip contains speech that extends well beyond the fallacy, potentially diluting the acoustic signal that is distinctive of the fallacious utterance. To quantify this effect, we compute, for each of the 1,891 samples in MM-USED-fallacy, the *character-level coverage ratio*: the number of characters in the original

fallacy span (Snippet) divided by the number of characters in the reference sentence(s) it was mapped to (SentenceSnippet). A ratio close to 100% indicates that the span roughly coincides with the full sentence, whereas a low ratio signals that a short fallacious expression was mapped to a much longer sentence.

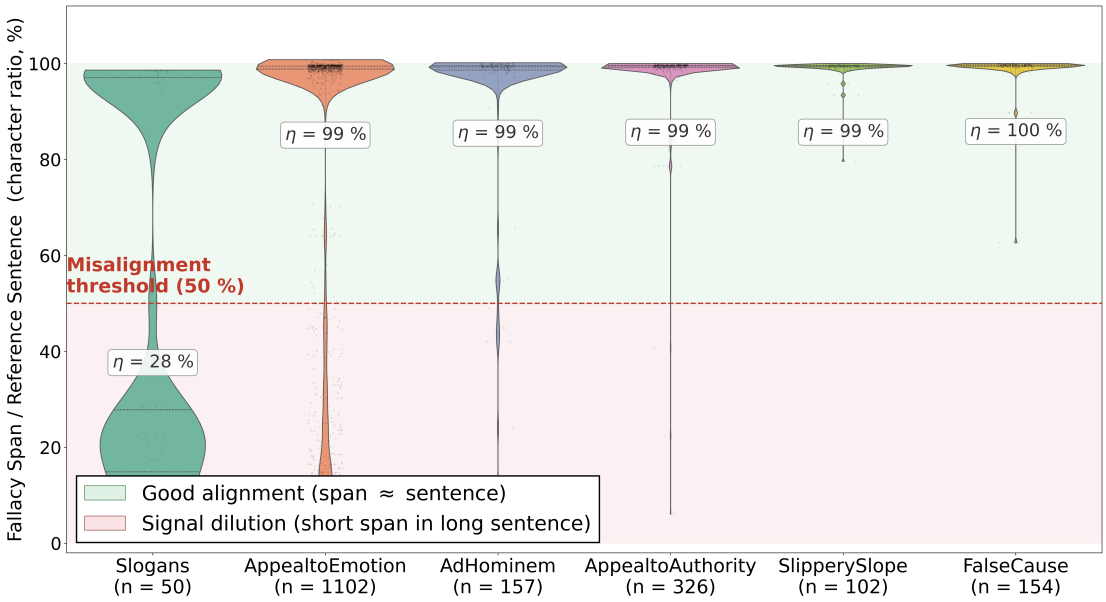


Figure 3.2: *Distribution of the character-level coverage ratio (fallacy span length / reference sentence length) per fallacy category, computed on the 1,891 samples in MM-USED-fallacy. The green zone indicates samples where the span closely matches the sentence boundaries; the red zone highlights cases where a short span was mapped to a much longer sentence, potentially diluting the audio signal. Each violin is annotated with η, the median of the coverage ratio for that category, i.e. the value below which half of the samples fall: a low η indicates that, for a typical sample in that category, the fallacy span covers only a small portion of the reference sentence.*

Figure 3.2 shows the distribution of coverage ratios per fallacy category. Overall, the median coverage is high (99.2%), confirming that the majority of samples are well aligned. However, 323 samples (17.1%) fall below a 50% coverage threshold, and this misalignment is *not* uniformly distributed across categories. *Slogans* and *Appeal to Emotion* are the most affected, with 60.0% and 25.5% of their samples below the threshold, respectively—even though *Appeal to Emotion* maintains a high median coverage ($\eta = 99\%$), a substantial tail of misaligned samples exists within the category. In contrast, *False Cause* and *Slippery Slope* exhibit near-perfect coverage (0% below the threshold), reflecting the fact that these fallacy types tend to span entire sentences or more.

Although our classification experiments target only the six macro fallacy cate-

gories, the subcategory annotations available in USED-fallacy allows us to perform a finer-grained diagnostic analysis of this coverage gap. Exploiting these annotations reveals that the misalignment within *Appeal to Emotion* is almost entirely concentrated in the *Loaded Language* subcategory, where 262 out of 575 samples (45.6%) fall below the 50% threshold, by far the highest misalignment rate among all subcategories. This is not surprising: loaded language by definition consists of individual emotionally charged words or short phrases embedded within longer utterances. Similarly, *Slogans* are inherently brief, catchy expressions that rarely fill an entire sentence. This pattern is not a data artefact but reflects the *linguistic structure* of these fallacy types: loaded language and slogans are designed to be concise and impactful, and their brevity is precisely what makes sentence-level audio alignment challenging for these specific categories and subcategories.

3.4.5 Methodology and Experimental Design

Task Formulation and Evaluation Protocol

We frame AFC as a multi-class sentence classification task. We evaluate models via leave-one-out cross-validation, totaling 28 individual model runs. In particular, we build folds such that, at each iteration, each debate is either in the left-out test split or in the remaining splits.

To address main fallacy categories only, leaving sub-categories for future work. To limit model overfitting on certain fallacy categories, we remove duplicate snippets, totaling 1063 unique dialogue-snippet pairs. To assess the contribution of each modality, we consider three input configurations: *text-only*, *audio-only*, and *text-audio*.

Architectural Approach

We experiment with the multimodal architecture presented in Section 3.3 (Figure 3.1). The text module comprises a pre-trained text embedding model and a dropout layer on top of it. The audio module consists of a pre-trained audio embedding model, a BiLSTM layer, and a dropout layer. The output of the text and audio modules is concatenated and fed to the classification module, defined as a stack of dense layers.

We extend the architectural framework established in Section 3.3 by exploring two audio signal encoding and three text encoding models. For audio, we consider Wav2Vec [289] and CLAP [363], while for text, we use pre-trained BERT [70],

RoBERTa [179], and SBERT [274]. We also experimented with pre-trained text embeddings without fine-tuning but obtained unsatisfactory results (not reported).

Training Configuration

We train all models using cross-entropy loss and Adam [142] for optimization. See Appendix B.5 for more details.

3.4.6 Results and Analysis

Table 3.14 reports the macro F1-score for each fallacy category, averaged across all model runs. The text-audio setting leads to performance improvement over text-only (up to 8 points) for BERT ($p\text{-}value<0.05$) and RoBERTa ($p\text{-}value>0.05$), independently of the choice of audio model. In contrast, we observe a significant performance drop for SBERT when adding audio.

The observed performance improvements are not equally distributed over fallacy types. For instance, text-only models achieve superior or equal F1-score for the Slogans (S) category. One reason for that could be the low number of examples in this category. Another reason could be that, according to linguistic analyses, slogans primarily rely on figurative language [79], whereas prosodic features have minimal impact on them [304].

Classifying fallacies at the sentence level may represent an additional challenge for the audio modality. As shown in Table 3.13 and Figure 3.2, a notable amount of samples in categories like *Appeal to Emotion* and *Slogans* are annotated at the span level in USED-fallacy, resulting in sentence-level audio clips that may not faithfully represent the fallacious utterance. However, the uniformly low audio-only scores across all categories, including those with near-perfect span-to-sentence coverage such as *False Cause* and *Slippery Slope*, suggest that audio signal dilution is not the primary factor limiting audio-only performance. Rather, audio features alone appear insufficient to discriminate between fallacy types, which likely explains why the audio modality is most effective when combined with text. Determining how to best leverage audio-text information remains an open challenge, as the introduction of audio induces variations across different classes, necessitating further studies.

Our investigation into the sensitivity of our results to individual speakers and the influence of audio features on fallacy classification revealed insightful nuances

	AE	AA	AH	FC	SS	S	Avg ($\bar{x} \pm \sigma$)
Text-only							
BERT	.70	.45	.15	.28	.22	.06	.32 ± .13
RoBERTa	.53	.50	.32	.29	.30	.17	.38 ± .18
SBERT	.54	.39	.23	.27	.20	.04	.31 ± .18
Audio-only							
Wav2Vec	.45	.05	.06	.08	.08	.03	.13 ± .07
CLAP	.29	.17	.03	.03	.09	.00	.12 ± .08
Text-Audio							
BERT + Wav2Vec	.80	.50	.13	.35	.23	.04	<u>.40 ± .17</u>
BERT + CLAP	.77	.44	.11	.31	.28	.01	.36 ± .17
RoBERTa + Wav2Vec	.70	.44	.16	.41	.30	.12	.39 ± .19
RoBERTa + CLAP	.74	.45	.23	.37	.31	.12	**.40 ± .19**
SBERT + Wav2Vec	.45	.29	.27	.21	.11	.04	.23 ± .11
SBERT + CLAP	.44	.32	.20	.25	.17	.04	.24 ± .10
Baselines							
Majority	.79	.00	.00	.00	.00	.00	.20 ± .17
Random	.33	.15	.08	.05	.03	.03	.12 ± .05

Table 3.14: *Result on MM-USED-fallacy. AE: Appeal to Emotion, AA: Appeal to Authority, AH: Ad Hominem, FC: False Cause, SS: Slippery Slope, S: Slogans. In bold the best model, underlined the second-best model.*

across different debates. While our dataset structure, comprising pairs of candidates per debate, prevented us from pinpointing specific speakers benefiting most from audio features, we observed significant variations across folds and debates. Notably, recent debates tended to show a more pronounced benefit from audio features, possibly due to improved audio quality or the predominance of newer audio files in the training data for audio feature extractors. Specifically, debates such as *1980-Carter/Reagan*, *2004-Bush/Kerry*, and *2008-McCain/Obama* notably benefited from the inclusion of audio features. In other debates, like *1976-Carter/Ford* and *1996-Clinton/Dole*, integrating audio features led to decreased classification accuracy, while debates like *1960-Kennedy/Nixon* yielded mixed results. These findings underscore the complex interplay between speaker characteristics, debate context, and the utility of audio features in enhancing fallacy classification, highlighting the need for further investigation.

3.4.7 Discussion and Findings

We posit that AFC should be framed as a multimodal task. To empirically evaluate our hypothesis, we build the first dataset for multimodal argumentative fallacy detection, MM-USED-fallacy. Our results show that the integration of audio modality is indeed beneficial, observing significant performance improvement (4-8 F1-score percentage points) in a variety of model architectures.

Our findings are coherent with recent studies in argument mining, including our foundational work presented in Section 3.3, as well as other research in multimodal argument mining [206] and fake news detection [121]. However, it is important to acknowledge that this study is based on a single dataset, and not all the text in USED-fallacy could be used, since in some cases audio-to-text alignment was unsuccessful. For this reason, some of the fallacies annotated in USED-fallacy are lost in MM-USED-fallacy.

Furthermore, while we argue that fallacy classification should be framed as a multimodal task, the annotations utilized in this research were derived from those defined in USED-fallacy based on textual content only. Such annotations are likely to disregard the potential insights coming from the acoustic elements of the debates. We believe that a multimodal formulation should affect the experimental setting starting from data collection, if possible, so as to capture several audio properties like non-verbal features [144] in addition to prosodic ones. In order to take into account all sound-related cues, a new annotation of fallacies should ideally be carried out from scratch, using a new set of guidelines specifically designed for multimodal analysis.

While this work focuses on AFC using a limited set of models and embedding approaches, it establishes the foundation for broader multimodal experimentation. The following sections build upon this resource, introducing an open toolkit and shared task that extend the investigation to new architectures, embedding strategies, and evaluation settings.

The multimodal resource that we provide thus serves as a foundational step that enables these broader experimental investigations. Moreover, to gain a deeper understanding of our results, employing interpretability techniques designed to emphasise the significance of paralinguistic elements over linguistic elements in prediction may be beneficial.

3.5 MAMKit: A Comprehensive Multimodal Argument Mining Toolkit

This section presents our comprehensive toolkit for MAM [196], developed in collaboration with Federico Ruggeri, Stefano Colamonaco, Andrea Zecca, Samuele Marro, and Paolo Torroni.

Key Contribution: Building on the foundational work presented in Section 3.3 and the fallacy classification advances described in Section 3.4, we introduce MAMKit, an open-source PyTorch toolkit designed to standardise and accelerate research in MAM.[a] While our previous contributions established important datasets and demonstrated the potential of multimodal approaches in political debate analysis, they also revealed significant challenges in audio encoding strategies and multimodal fusion techniques. MAMKit addresses these limitations by providing a unified platform that facilitates reproducible research, systematic comparison of approaches, and rapid integration of new contributions to the field.

[a] Code available at `https://github.com/nlp-unibo/mamkit`

The toolkit encompasses four datasets and six distinct model architectures, including three novel fusion strategies we introduce: CSA (Concatenation-based early fusion with Self-Attention), Ensemble (weighted late fusion), and Mul-TA (crossmodal attention). Through comprehensive evaluation across all supported datasets and tasks, we provide insights into the current state and future directions of MAM research, while establishing a foundation for community-driven advancement in this emerging field.

MAMKit is an open-source, publicly available PyTorch toolkit designed to access and develop datasets, models, and benchmarks for MAM. It provides a flexible interface for accessing and integrating datasets, models, and preprocessing strategies through composition or custom definition. MAMKit is designed to be extendible, ensure replicability, and provide a shared interface as a common foundation for experimentation in the field. At the time of writing, MAMKit offers 4 datasets and 6 distinct model architectures, along with audio and text processing capabilities, organised in 5 main components (see Figure 3.3).

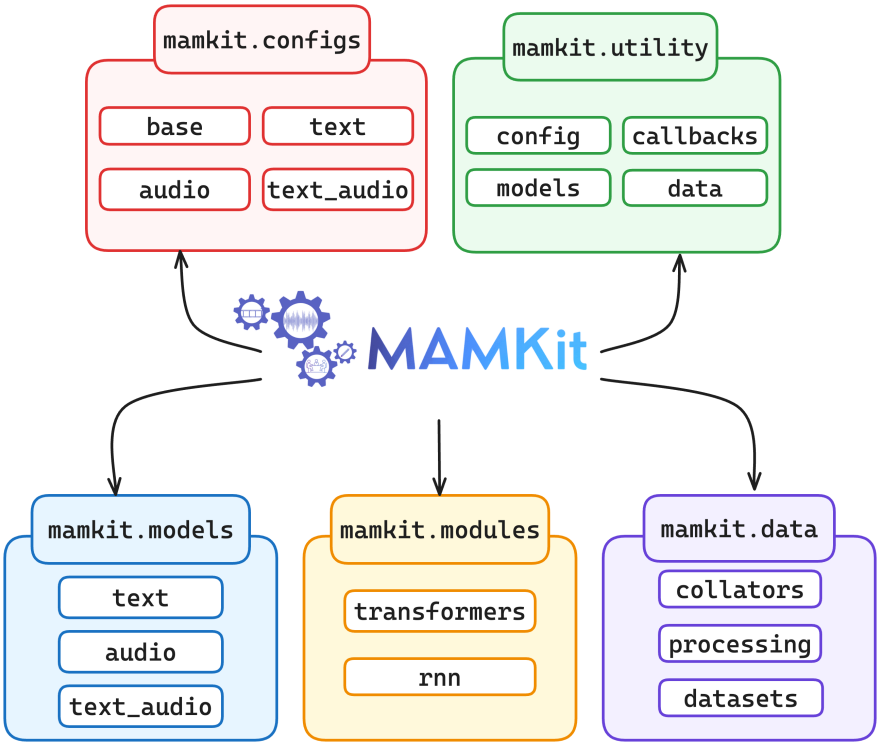


Figure 3.3: *Overall architecture of MAMKit.*

3.5.1 MAMKit

Description of toolkit components

MAMKit is structured into five core packages that together provide a complete pipeline for multimodal argument mining research. The `mamkit.data` package handles dataset loading, preprocessing, and batch preparation. The `mamkit.models` package contains model architectures for audio-only, text-only, and multimodal configurations. The `mamkit.modules` package provides shared neural network components used across different models. The `mamkit.utility` package offers supporting functionality for data management, model training, and evaluation. Finally, the `mamkit.configs` package enables easy instantiation of models through predefined configuration keys, facilitating reproducibility and benchmarking. Below, we describe each component in detail.

Datasets — The `mamkit.data` package covers dataset creation (`data.datasets`) and preprocessing (`data.preprocessing` and `data.collators`). The `data.datasets` module provides the `Loader` interface, a general-purpose wrapper for datasets, covering data downloading, task-specific data parsing, and data inter-

facing. Regarding the latter functionality, the module includes ad-hoc implementations for unimodal (`UnimodalDataset`) and multimodal (`MultimodalDataset`) data based on the PyTorch `Dataset` interface. The `data.processing` module provides the `Processor` interface for defining custom data processing and implements unimodal (`UnimodalProcessor`) and multimodal (`MultimodalProcessor`) processing steps. For instance, the `AudioTransformer` class implements transformer-based audio processing. Similarly to `data.processing`, the `data.collators` module is designed to address input processing at the batch level, in compliance with PyTorch `DataLoader` APIs. The module includes implementations for unimodal (`UnimodalCollator`) and multimodal (`MultimodalCollator`) input batches.

Models — The `mamkit.models` package holds definitions for the supported models. It provides `models.audio`, `models.text` and `models.text_audio` modules. Each model implements the `torch.nn.Module` interface that can be extended to define the models for each input configuration.

Modules — The `mamkit.modules` package handles the definition of the shared model layers, such as `transformer_modules`.

Utility — The `mamkit.utility` package contains classes and methods used by other modules. For example, the `utility.data` module contains methods for downloading data from web storages or GitHub repositories, while `utility.model` manages the overall training and evaluation lifecycles. Currently, the `MAMKitLightningModel` class and the `to_lightning_model()` method are used to wrap models as PyTorch Lightning [86] models to leverage their functionalities for training and evaluation. Incorporating PyTorch Lightning in our toolkit streamlines training and evaluation with a simplified loop, standardized interface, reproducibility, performance optimizations, accelerator integration, logging capabilities, and extensive community support.

Configs — The `mamkit.configs` package serves as a streamlined interface for accessing model configurations across three modalities: audio, text, and text-audio. At its core, the `config.base` module establishes two fundamental classes: `ConfigKey`, defining configuration keys, and `BaseConfig`, providing a base configuration structure. This architecture simplifies benchmarking efforts by enabling users to instantiate models via designated configuration keys. Consequently, leveraging models with exact parameter setups for benchmarking or further experimen-

tation becomes straightforward, enhancing research reproducibility and efficiency within the toolkit.

Example Usage

MAMKit's design facilitates access to existing datasets and models and supports future development. In this section, we present several examples to illustrate common use cases.

Data Loading — An important feature of MAMKit is its unified and straightforward interface for data access. Several MAM datasets are included in MAMKit. Adding a new dataset to MAMKit requires defining a new subclass of `Loader`, extending it with the specific information needed to access and reconstruct the dataset. In the example that follows, a dataset is loaded using the `MMUSED` class from `mamkit.data.datasets`, which extends the `Loader` interface and implements specific functionalities for data loading and retrieval. Users can specify task and input mode (`text-only`, `audio-only`, or `text-audio`) when loading the data, with options to use default splits or load splits from previous works. The example uses splits from our foundational work.

```
from mamkit.data.datasets import UKDebates, InputMode
loader = UKDebates(
task_name='asd',
input_mode=InputMode.TEXT_ONLY,
base_data_path=base_data_path)
split_info = loader.get_splits('foundational-work')
```

The `get_splits` method of the `loader` returns data splits in the form of a `data.datasets.SplitInfo`. The latter wraps split-specific data, each implementing Pytorch's Dataset interface and compliant to the specified input modality (i.e., `text-only`). The `Loader` interface also allows users to integrate methods defining custom splits as follows:

```
from mamkit.data.datasets import SplitInfo
def custom_splits(self) -> List[SplitInfo]:
train_df = self.data.iloc[:50]
val_df = self.data.iloc[50:100]
test_df = self.data.iloc[100:]
fold_info = self.build_info_from_splits(train_df=...)
return [fold_info]
loader.add_splits(method=custom_splits,
key='custom')
```

```
split_info = loader.get_splits('custom')
```

Modelling — MAMKit offers a simple method for defining custom models and leveraging models from the literature. Utilising the same interface for both tasks aims to simplify access to existing models and establish new ones with a coherent structure. This will hopefully facilitate the spread of established models and encourage the development of new ones by maintaining consistency throughout the process. The example below illustrates that defining a custom model is straightforward. It entails creating the model within the `models` package, specifically by extending either the `AudioOnlyModel`, `TextOnlyModel`, or `TextAudioModel` classes in the `models.audio`, `models.text`, or `models.text_audio` modules, respectively, depending on the input modality handled by the model.

```
from mamkit.models.text import Transformer
model = Transformer(
model_card='bert-base-uncased',
dropout_rate=0.1, ...)
```

The following example demonstrates how to instantiate a model with a configuration found in the literature. This configuration is identified by a key, `ConfigKey`, containing all the defining information. The key is used to fetch the precise configuration of the model from the `configs` package. Subsequently, the model is retrieved from the `models` package and configured with the specific parameters outlined in the configuration.

```
from mamkit.configs.base import ConfigKey
from mamkit.configs.text import TransformerConfig
from mamkit.data.datasets import InputMode
config_key = ConfigKey(
dataset='mmused',
task_name='asd',
input_mode=InputMode.TEXT_ONLY,
tags={'foundational-work'})
config = TransformerConfig.from_config(
key=config_key)
model = Transformer(
model_card=config.model_card,
dropout_rate=config.dropout_rate
...)
```

In both the described use cases, the model is then encapsulated into a Pytorch Lightning model, and training and evaluation are conducted by leveraging the methods provided by this wrapper.

```
from mamkit.utility.model import to_lighting_model
import lightning
model = to_lighting_model(model=model,
num_classes=config.num_classes,
loss_function=...,
optimizer_class=...)
trainer = lightning.Trainer(max_epochs=100,
accelerator='gpu',
...)
trainer.fit(model,
train_dataloaders=train_dataloader,
val_dataloaders=val_dataloader)
```

Benchmarking — The `mamkit.configs` package simplifies reproducing literature results in a structured manner. Upon loading the dataset, experiment-specific configurations can be easily retrieved via a configuration key. Specifically, unlike the examples reported in Section 3.5.1, where configurations refer just to a model implementation, in the below example, they encompass both data processing and model parameterisation based on previous literature work. This enables instantiating a processor using the same features processor employed in the experiment. In the example below, we adopt a configuration akin to our foundational work in Section 3.3, employing a BiLSTM model with audio encoded with MFCC features. Hence, we define a `MFCCExtractor` processor using configuration parameters. Data splits are loaded using the experiment reference key, mirroring what was shown in Section 3.5.1.

```
from mamkit.configs.audio import BiLSTMMFCCsConfig
from mamkit.configs.base import ConfigKey
from mamkit.data.datasets import UKDebates, InputMode
from mamkit.data.processing import MFCCExtractor, UnimodalProcessor
from mamkit.models.audio import BiLSTM
loader = UKDebates(task_name='asd',
input_mode=InputMode.AUDIO_ONLY)
config = BiLSTMMFCCsConfig.from_config(
key=ConfigKey(dataset='ukdebates',
input_mode=InputMode.AUDIO_ONLY,
task_name='asd',
tags='foundational-work'))
for split_info in loader.get_splits(
key='foundational-work'):
processor =
UnimodalProcessor(
```

```
features_processor=MFCCExtractor(
mfccs=config.mfccs, ...))
split_info.train = processor(split_info.train)
...
model = BiLSTM(embedding_dim=
                config.embedding_dim, ...)
```

Models

MAMKit comes with 3 models from the MAM literature and 3 original models we contribute based on state-of-the-art unimodal audio encoders and fusion strategies. All models comply with the following architecture: text and audio modules for encoding individual modalities, a fusion layer to merge them, and a final classification head tailored to the downstream task of interest. Table 3.15 provides a summary.

Model	Text Encoding	Audio Encoding	Fusion
BiLSTM [197]	GloVe + BiLSTM	(Wav2Vec2 ∨ MFCCs) + BiLSTM	Concat-Late
MM-BERT [198]	BERT	(Wav2Vec2 ∨ HuBERT ∨ WavLM) + BiLSTM	Concat-Late
MM-RoBERTa [198]	RoBERTa	(Wav2Vec2 ∨ HuBERT ∨ WavLM) + BiLSTM	Concat-Late
CSA (Ours)	BERT	(Wav2Vec2 ∨ HuBERT ∨ WavLM) + Transformer	Concat-Early
Ensemble (Ours)	BERT	(Wav2Vec2 ∨ HuBERT ∨ WavLM) + Transformer	Avg-Late
Mul-TA (Ours)	BERT	(Wav2Vec2 ∨ HuBERT ∨ WavLM) + Transformer	Cross

Table 3.15: *Multimodal models available in MAMKit. For each model, we summarize its text and audio encoding modules and its fusion strategy. Concat: Concatenation; Avg: Average; Cross: Crossmodal Attention.*

We refer to the fusion strategies as follows:

- **Concatenation**: combines features (*early fusion*) or embeddings from single modality architectures (*late fusion*) from all modalities into a single vector by concatenating them;
- **Average**: merges features (*early fusion*) or embeddings from single modality architectures (*late fusion*) by simply averaging information from each modality;
- **Crossmodal Attention**: attends to interactions between multimodal sequences across distinct time steps and facilitates the transfer of streams from one modality to another.

BiLSTM [197] — The text module comprises a pre-trained GloVe [250] embedding layer and a stack of BiLSTM layers. Similarly, the audio module is a stack

of BiLSTM layers. The fusion strategy is vector concatenation. The classification head is a MLP.

MM-BERT, MM-RoBERTa [198] — The text module comprises a trainable text embedding model and a dropout layer on top. The audio module comprises a pre-trained audio embedding model, a BiLSTM layer, and a dropout layer. The output of the text and audio modules is concatenated and fed to the classification module, defined as a stack of dense layers.

CSA (Ours) — A multimodal transformer inspired by Yu et al.[376], whereby text and audio embeddings are concatenated along the time dimension, and a self-attention layer is employed to obtain a cross-modal text and audio embedding. This embedding is averaged over time and fed to a classification head.

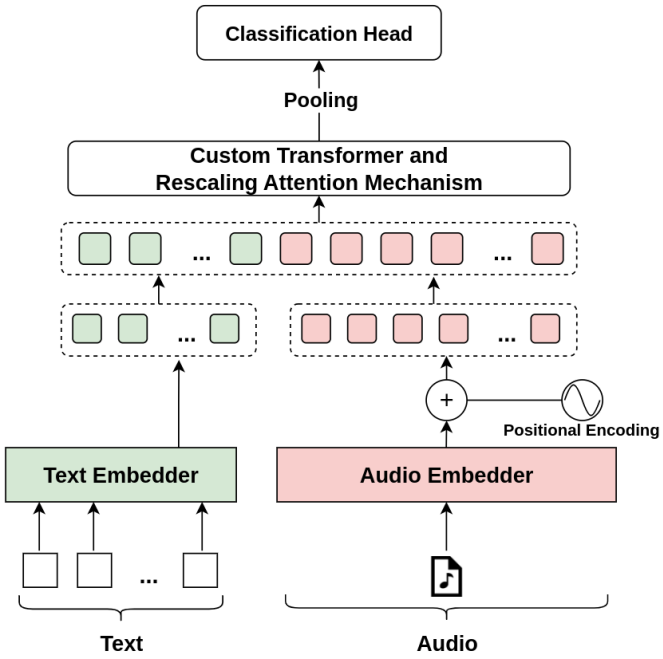


Figure 3.4: *The CSA model architecture.*

The main issue of this architecture is the significant difference between the lengths of the audio and text sequences. Even with downscaling, the audio embeddings tend to be significantly longer (often by a factor of ∼10). Consequently, audio features dominate the early stages of training, leading to underwhelming performance. To address this issue, we develop a novel transformer variant in which we reweight the attention scores of text and audio sequences for each layer. Let m be the length of the text sequence and n the length of the audio sequence. We rescale the attention scores of the text sequence by $\frac{m+n}{2m}$ and of the audio

sequence by $\frac{m+n}{2n}$. This reweighting ensures that text and audio sequences have the same total weight. This reweighting ensures that text and audio sequences have the same total weight. Figure 3.4 summarises our Concatenation-based early fusion with the Self-Attention (CSA) transformer model.

Ensemble (Ours) — This architecture consists of two independent unimodal models for text and audio, respectively. A weighted average of the probability vectors of the unimodal classification heads constitutes the final prediction.

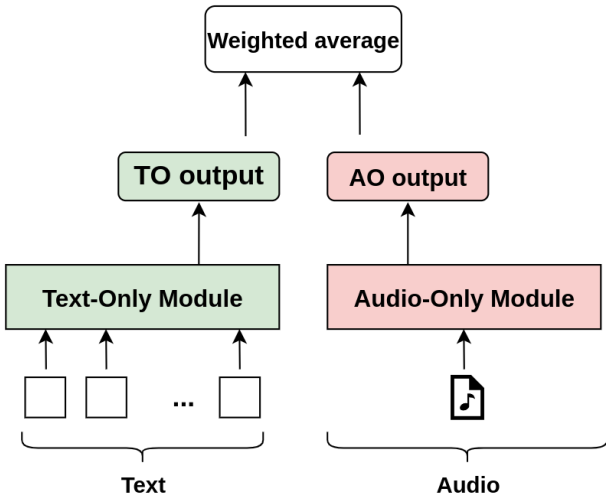


Figure 3.5: *The Ensemble model architecture.*

The text-only model involves averaging BERT embeddings along the time dimension and feeding them to a two-layer classification head. The audio-only model follows the same architecture as the text-only model, although with a custom transformer which is trained along with the head. The main challenge is determining how to merge the outputs of the two unimodal classification heads. We compute a weighted average with weight w_e defined as follows:

$$w_e = l + (u - l) \cdot \frac{\tanh w + 1}{2} \tag{3.1}$$

where w is a learnable parameter in the $[l, u]$ range. Bounding ensures that the ensemble is forced to exploit the output of both classification heads, preventing a *dead neuron* situation where the ensemble focuses on a single modality only. We set $l = 0.3$ and $u = 0.7$ for learning stability. Figure 3.5 summarises Ensemble.

Mul-TA (Ours) — We propose a variant of the MulT architecture [337]: a transformer model for carrying out multimodal tasks without the need for modality alignment. The core module of MulT is the directional pairwise cross-modal

attention layer, which captures interdependencies between multimodal sequences and seamlessly adjusts information flow between modalities. In practice, the cross-modal attention layer uses one modality A as the query vector and another modality B as key and value vectors. The layer is applied for each pair of input modalities. Pairs with the same B modality are combined into a unified sequence using a self-attention layer. Lastly, each unified sequence is averaged over the time dimension and concatenated. The resulting embedding vector is fed to a classification head. While MulT was developed for three modalities, totalling six (A, B) pairs, our variant uses only two, totalling two (A, B) pairs. Additionally, we replace the self-attention unification step with an average.

Figure 3.6 summarises Mul-TA, our MulT architecture variant, tailored to text and audio modality.

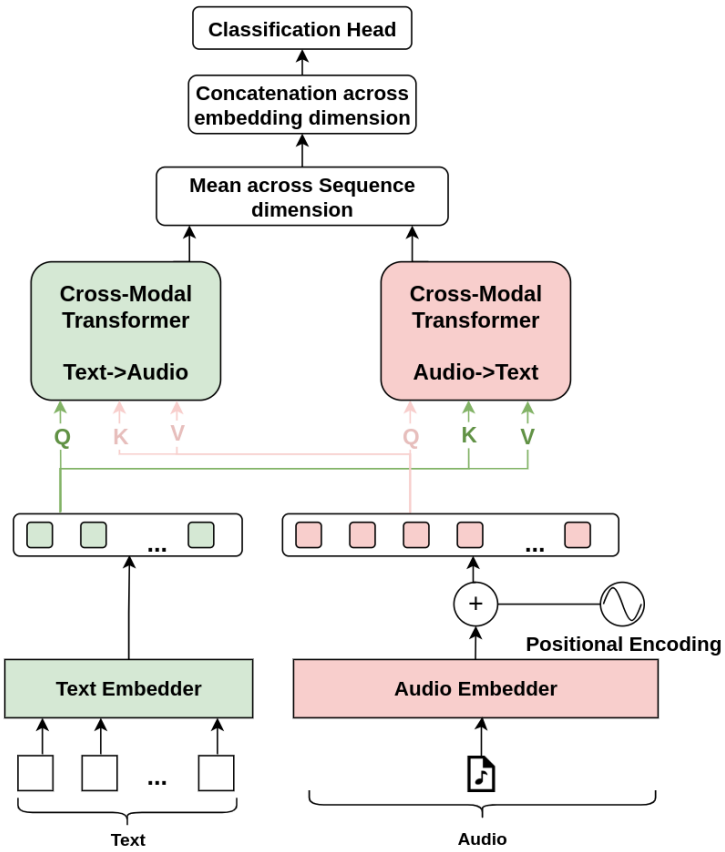


Figure 3.6: *The Mul-TA model architecture.*

Data

Four datasets are currently available within MAMKit: UKDebates [177], M-Arg [207], MM-USED [197], and MM-USED-fallacy [198]. We present comprehensive descriptions of the first three datasets in Section 3.3 and of MM-USED-fallacy in Section 3.4.

Tasks

The tasks currently supported by MAMKit are derived from literature [178, 163], and comprise ASD, ACC, ARC and AFC.

3.5.2 Experiments

We employ MAMKit to provide a robust and reproducible overview of a significant share of the work published on MAM so far. In particular, we evaluate MAMKit supported models on all available tasks and datasets. We build our evaluation as follows. Regarding model evaluation, we compute macro F1-score except on UKDebates, for which we report binary F1-score [177]. We carry out a repeated five-fold cross-validation routine for UKDebates and M-Arg$_\gamma$ using the same folds defined using the same folds defined in our foundational work (Section 3.3). Similarly, we perform a repeated train and test routine for MM-USED on official data splits [103]. We set the number of repetitions to three in both cases. Lastly, we perform a leave-one-out routine for MM-USED-fallacy following our multimodal fallacy classification work in Section 3.4. See Appendix C.1 for additional details.

3.5.3 Results

Table 3.16 reports the best classification performance for each model. See Appendix C.2 for all results.

UKDebates — We observe no notable benefits in integrating the audio modality in all models, comparable to the results from our foundational work (Section 3.3). Specifically, multimodal models show equal or lesser classification performance than their text-only modules.

M-Arg$_\gamma$ — Our results significantly differ from those reported in Section 3.3. In particular, Ensemble and Mul-TA are noticeably underperforming compared to their text-only counterparts. The only exceptions are MM-BERT and CSA

Model	UKDebates (ASD)	M-Arg$_\gamma$ (ARC)	MM-USED (ASD)	MM-USED (ACC)	MM-USED-fallacy (AFC)
			Text Only		
BiLSTM (T_1)	$.552_{\pm.047}$	$.120_{\pm.006}$	$.811_{\pm.004}$	$.663_{\pm.002}$	$.525_{\pm.113}$
BERT (T_2)	$.654_{\pm.003}$	$.132_{\pm.004}$	$.824_{\pm.009}$	$.679_{\pm.004}$	$.594_{\pm.122}$
RoBERTa (T_3)	$.692_{\pm.005}$	$.172_{\pm.015}$	$.839_{\pm.010}$	$.680_{\pm.001}$	$.615_{\pm.097}$
			Audio Only		
BiLSTM (A_1)	$.393_{\pm.040}$	$.024_{\pm.012}$	$.774_{\pm.008}$	$.596_{\pm.005}$	$.657_{\pm.000}$
Transformer (A_2)	$.455_{\pm.004}$	$.000_{\pm.000}$	$.771_{\pm.019}$	$.526_{\pm.004}$	$.629_{\pm.162}$
			Text Audio		
BiLSTM ($T_1 + A_1$)	$.533_{\pm.009}$	$.084_{\pm.016}$	$.815_{\pm.006}$	$.667_{\pm.000}$	$.572_{\pm.099}$
MM-BERT ($T_2 + A_1$)	$.662_{\pm.004}$	$.160_{\pm.015}$	$.841_{\pm.005}$	$.680_{\pm.004}$	$.599_{\pm.128}$
MM-RoBERTa ($T_3 + A_1$)	$.687_{\pm.010}$	$.178_{\pm.012}$	$.837_{\pm.009}$	$.678_{\pm.003}$	$.624_{\pm.074}$
CSA ($T_2 + A_2$)	$.663_{\pm.014}$	$.160_{\pm.015}$	$.833_{\pm.011}$	$.693_{\pm.001}$	$.582_{\pm.114}$
Ensemble ($T_2 + A_2$)	$.586_{\pm.015}$	$.011_{\pm.011}$	$.826_{\pm.011}$	$.681_{\pm.002}$	$.612_{\pm.134}$
Mul-TA ($T_2 + A_2$)	$.616_{\pm.019}$	$.098_{\pm.031}$	$.837_{\pm.006}$	$.697_{\pm.003}$	$.605_{\pm.110}$

Table 3.16: *Test classification performance on MAM datasets. For each multimodal model, we report its constituting text module (T_i) and audio module (A_j).*

with slightly higher performance. Additionally, audio-only models fail to learn the task.

MM-USED — We observe a small performance gap between audio-only and text-only models, suggesting that the audio modality may be a valuable indicator in both ASD and ACC tasks. However, multimodal models achieve comparable performance to their text-only counterparts, with minor improvements only for MM-BERT (+1.7), CSA (+0.9), Ensemble (+0.2) and Mul-TA (+1.3) in ASD, CSA (+1.4), and Mul-TA (+1.8) in ACC.

MM-USED-fallacy — In contrast to other tasks and datasets, in MM-USED-fallacy, audio-only models are the best-performing ones. The performance of text-audio models is slightly better than that of the corresponding text-only models but below that of audio-only models. Alternative fusion strategies yield only a moderate, non-systematic improvement.

3.5.4 Discussion and Findings

MAM is an emerging and largely unexplored research domain with significant potential for theoretical and applied innovation. At this stage, building a collaborative and standardised platform for experimentation is essential to strengthen the research community and accelerate progress. To this end, we developed MAMKit, an open-source PyTorch toolkit that provides access to multiple datasets, state-of-the-art models, and processing strategies, offering a unified environment for

benchmarking and experimentation in multimodal argument mining.

Our experiments show that the advanced audio encoding and fusion techniques implemented in MAMKit do not yet produce the expected performance improvements. This limitation may stem from the current model architectures, suggesting the need for further research on more expressive encoding and fusion mechanisms. Another likely factor is the annotation process of existing datasets, where labels were created on textual transcripts and later extended to audio, without considering paralinguistic features during annotation. As discussed in Section 3.4, this constrains the capacity of multimodal architectures to exploit acoustic cues fully. Revising existing corpora to include auditory information natively would be a necessary step toward realising the full potential of MAM.

While MAMKit represents a substantial step forward for the field, several limitations remain. The toolkit currently supports only PyTorch, which may limit its adoption by researchers using other frameworks such as TensorFlow. In addition, not all datasets and models relevant to multimodal argument mining are yet integrated. Resources such as VivesDebate-Speech [282], ImageArg [181], and MMClaims [43], as well as models like M-ArgNet [207], are not currently implemented. Moreover, the present version of MAMKit focuses exclusively on text and audio, underlining the importance of future extensions toward visual modalities.

Looking ahead, MAMKit offers a foundation for expanding multimodal argument mining to new modalities and tasks. Planned extensions include integrating the ImageArg dataset, which supports stance and image persuasiveness classification and was further expanded during the ImageArg-2023 shared task [180], as well as incorporating MMClaims for multimodal claim detection in social media. These additions will support the exploration of vision-language argument mining and facilitate transfer learning across diverse domains. Ultimately, advancing multimodal discourse analysis will require close collaboration between the argumentation and signal-processing communities, and MAMKit aims to serve as a shared platform for fostering such interdisciplinary research.

3.6 Overview of MM-ArgFallacy2025 on Multimodal Argumentative Fallacy Detection and Classification in Political Debates

This section presents MM-ArgFallacy2025 [199], the first shared task dedicated to multimodal argumentative fallacy detection and classification, organised in collaboration with Federico Ruggeri, Serena Villata and Paolo Torroni.

Key Contribution: Building on the toolkit infrastructure developed in Section 3.5 and the multimodal fallacy classification advances presented in Section 3.4, we designed a community-wide evaluation that addresses fundamental challenges in MAM while providing standardised benchmarks for the field.[a] The shared task was motivated by the need to systematically evaluate different approaches to multimodal fallacy analysis and to foster broader community engagement in this emerging research area. While our previous work demonstrated the potential of multimodal approaches for fallacy classification, it also revealed significant challenges in effectively leveraging audio information and achieving robust multimodal fusion. MM-ArgFallacy2025 was designed to address these challenges through structured competition across both binary detection and multi-class classification tasks, enabling a comprehensive comparison of methodological approaches while expanding the available evaluation data.

[a]Official shared task website: `https://nlp-unibo.github.io/mm-argfallacy/2025/`

The shared task attracted multiple international teams and established important baselines for the field. Through systematic evaluation across textual, audio, and multimodal approaches, the competition revealed both the current capabilities and fundamental limitations of existing methods, providing crucial insights for future research directions in multimodal argumentation analysis.

3.6.1 Problem Formulation

MM-ArgFallacy2025's subtasks are formulated as follows.

Argumentative Fallacy Detection (AFD) — The input is a sentence, in the form of text or audio or both, extracted from a political debate. The objective is

to determine whether the input contains an argumentative fallacy.

AFC — The input is a sentence, in the form of text or audio or both, extracted from a political debate, containing a fallacy. The objective is to determine the type of fallacy contained in the input, according to the classification introduced by [94].[5] In particular, the fallacy categories are: Appeal to Emotion, Appeal to Authority, Ad Hominem, False Cause, Slippery Slope, and Slogan. Examples of each fallacy category were provided in Table 3.10 in Section 3.4. For each subtask, participants can leverage the debate context of a given input: all its previous sentences and corresponding aligned audio samples. For instance, consider the **text-only** input setting. Given a sentence from a political debate at index i, participants can use sentences with indexes from 0 to $i-1$, where 0 denotes the first sentence in the debate.

3.6.2 Data

We describe the available training data for the challenge and the data collection process to curate the test set used to evaluate participants' submissions in the challenge (hereinafter, denoted as *secret test set*). All datasets are made available through MAMKit [196][6]. Since most of these multimodal datasets cannot release audio samples for copyright reasons, MAMKit provides a simple interface to dynamically build them and foster reproducible research.

Training Data

The primary training dataset is MM-USED-fallacy [198], developed in our work presented in Section3.4. The data provides annotations for AFC and AFD subtasks. Additionally, inspired by [94]'s observations on the benefits of employing other AM tasks like component detection for fallacy detection and classification, participants could use UKDebates [177], M-Arg [207], MM-USED [197] to encourage multi-task training approaches (see Table 3.17 for a summary). Detailed descriptions of UKDebates, M-Arg, and MM-USED are provided in Section 3.3, while MM-USED-fallacy is described in Section 3.4.

Text to Audio Alignment Corrections — Compared to the initial release of MAMKit [196], we introduce an improved text-to-audio alignment framework based on WhisperX [14], a state-of-the-art speech recognition model that allows

[5] We only refer to macro categories while sub-categories are left for future work.
[6] `https://nlp-unibo.github.io/mamkit/`

Name	No. Samples	Task$^{\alpha}$	Domain
Primary			
MM-USED-fallacy	18,925; 3,338	AFD; AFC	US Presidential Elections
Supplementary			
UKDebates	386	ASD	UK Prime Ministerial Elections
M-ARG	4,104 / 2,443	ARC	US Presidential Elections
MM-USED	26,781	ASD, ACC	US Presidential Elections

Table 3.17: *Available datasets for MM-ArgFallacy2025 shared task.*
$^{\alpha}$*Following [196], we denote tasks as Argumentative Fallacy Detection (AFD), Argumentative Fallacy Classification (AFC), Argumentative Sentence Detection (ASD), Argumentative Relation Classification (ARC), and Argumentative Component Classification (ACC).*

for precise and fine-grained audio-to-text alignment. We use this framework to address some well-known alignment issues in MM-USED [197] and MM-USED-fallacy [198], allowing us to integrate previously discarded debates and favouring the collecting of novel data. The updated datasets are available through MAMKit.

Test Data

Data Collection — We collect and annotate novel debates from US politics available in The American Presidency Project[7]. In particular, we consider the first presidential debate of the election cycle in Atlanta between Trump and Biden, aired on 28th June 2024, and the first presidential debate between Trump and Harris, aired on 11 September 2024. We follow the data collection pipeline proposed in our foundational work in Section 3.3 to retrieve the original audio recordings, but improve the text-to-audio alignment by leveraging WhisperX [14] for transcription, alignment and diarization. We follow Goffredo et al. [94] and split debates into paragraphs, where each paragraph corresponds to a speaker's turn. We use our text-to-audio alignment framework to pair paragraphs and corresponding textual sentences with related audio chunks. In total, we obtained 134 paragraphs for the first debate and 163 for the second one.

Data Annotation — For the annotation, we instruct two annotators with expertise in AM tasks and near-native English proficiency. We provide annotators with the guidelines of Goffredo et al.[94] for detecting and classifying argumentative fallacies. This is required to ensure annotation consistency with existing datasets

[7] `https://www.presidency.ucsb.edu/`

(e.g., MM-USED-fallacy). We remove paragraphs (i.e., dialogue turns) that do not belong to the main speakers of the debate (i.e., Trump, Biden and Harris). In total, we obtain 154 paragraphs for annotation, equivalent to 2154 sentences. We rely on Label Studio[8] for annotation, an open-source data annotation platform. For AFD and AFC subtasks, we provide annotators with the same instructions described in Section3.6.1. In particular, for AFD, annotators label each sentence in a given paragraph from a debate as containing a fallacy. In case an annotator labels a sentence as fallacious, they also provide the corresponding logical fallacy category to address AFC subtask.

IAA — Since fallacies can span multiple sentences [94], we report the rate of exact and partial overlaps between annotations. An exact overlap is when both annotators agree on all sentences constituting a fallacy. In contrast, a partial overlap is when annotators agree on a subset of sentences constituting a fallacy. We observe 236 overlaps, 110 of which are exact, while the remaining 126 are partial. The agreement rate, measured as the number of sentences detected as fallacious of the same category by both annotators, is 67.37%. Moreover, we compute IAA at the sentence level, measured as Cohen's Kappa [56]. For AFD, the IAA is 0.4787 (*moderate agreement*), while, for AFC, the IAA is 0.4954 (*moderate agreement*). Additionally, regarding AFC, the per-category IAA is as follows: 0.411 Appeal to Emotion, 0.337 Appeal to Authority, 0.357 Ad Hominem, 0.224 False Cause, and 0.712 Slogan. No annotator labelled a fallacy instance as Slippery Slope.

Fallacy Category	**No. Instances**
Appeal to Emotion	142
Appeal to Authority	16
Ad Hominem	46
False Cause	16
Slippery Slope	0
Slogan	9
Total	229

Table 3.18: *Secret test set statistics for Argumentative Fallacy Classification.*

Resulting Dataset — Table 3.18 reports the statistics of the resulting secret test set. We observe that a large majority of fallacies belong to Appeal to Emotion,

[8] `https://labelstud.io/`

followed by Ad Hominem. These findings are in line with annotations reported in previous work on argumentative fallacy classification [94]. Regarding AFD, the collected secret test set contains 229 fallacies and 1,946 non-fallacious sentences.

3.6.3 Overview of the Systems and Results

Three teams participated in the AFC subtask, while two teams participated in the AFD subtask. In total, participants submitted 25 valid runs. No team participated in both subtasks.

Team	Rank	Text	Rank	Audio	Rank	Text-Audio
		AFC				
Team NUST	1	**0.4856**	2	0.1588	1	**0.4611**
Team AlessioPittiglio	3	0.4444	1	**0.3559**	2	<u>0.4403</u>
Team CASS	5	0.1432	4	0.0864	5	0.1432
Baseline BiLSTM	2	<u>0.4721</u>	3	0.1582	4	0.2191
Baseline Transformer	4	0.3925	5	0.0643	3	0.3816
		AFD				
Team Ambali_Yashovardhan	2	<u>0.2534</u>	1	**0.2095**	3	0.2244
Team EvaAdriana	4	0.2195	2	<u>0.1690</u>	4	0.1931
Baseline BiLSTM	3	0.2462	3	0.0000	2	<u>0.2337</u>
Baseline Transformer	1	**0.2770**	3	0.0000	1	**0.2848**

Table 3.19: *Results for multimodal argumentative fallacy detection on political debates. For AFC, we report the macro F1-score, while for AFD, we report the binary F1-score. Best results per subtask are in **bold**, second best results are <u>underlined</u>.*

Table 3.19 shows the results achieved by the individual teams for each subtask. Regarding AFC, we observe that only one team, Team NUST [320], beats baselines on the text-only modality with a F1-score of 0.4856. This result shows that even a simple baseline like a BiLSTM is a strong competitor. In contrast, all participants improved over the baselines when considering the audio modality, while two teams surpassed the transformer baseline in the multimodal setting. Regarding AFD, Team Ambali_Yashovardhan reaches comparable performance to baselines in the text-only input setting, achieving rank 2. Nonetheless, despite reporting significant results in the audio-only setting where baselines fail the task, their solution is outperformed by both baselines in the multimodal setting.

All teams used neural networks, with transformer-based models being the most frequent choice. Some teams also employed machine learning classifiers like XGBoost on top of neural network models. Moreover, several teams explored a wide set of solutions to account for class imbalance, a well-known challenge in fallacy

detection [94].

Baselines — For both tasks, we employ the same set of baselines: a feature-based BiLSTM [198] and a transformer-based model [196]. Regarding the BiLSTM model, the baseline uses GloVe embeddings for text inputs and extracted MFCCs features for audio recordings. Conversely, the transformer-based model uses RoBERTa for encoding text and WavLM for audio. Both architectures employ a late fusion strategy for the multimodal setting, where text and audio embeddings are concatenated and fed to a final classification layer. Independent of the given input setting, we denote the baselines as Baseline BiLSTM and Baseline Transformer, respectively.

System Descriptions and Task-Specific Results — Below, we describe the approaches of all participating systems; see also Table 3.20 for an overview.

Team	Task		Text							Audio					MAM			Misc				
	AFC	AFD	BERT	RoBERTa	DeBERTa	ModernBERT	SBERT	ALBERT	DeepSeek R1	Whisper	Wav2Vec	WavLM	HuBERT	MFCCs	Late Fusion	Early Fusion	Mid Fusion	Data Augmentation	Class Weighting	Bayesian Optimization	Focal Loss	Context
NUST [320]	☑			☑						☑					☑			☑	☑			
AlessioPittiglio [253]	☑		☑	☑	☑	☑					☑	☑	☑		☑	☑	☑	☑	☑	☑		☑
CASS [133]	☑			☑							☑			☑	☑							
Ambali_Yashovardhan		☑		☑									☑		☑						☑	
EvaAdriana [162]		☑	☑	☑			☑	☑	☑		☑			☑	☑							

Table 3.20: *Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking.*

Argumentative Fallacy Classification. Team NUST [320] employ RoBERTa for text encoding and Whisper for audio encoding. The two encodings are combined in a late fusion fashion without requiring joint end-to-end training and fed to an XGBoost [49] classifier. To account for label imbalance, they propose several solutions, including generating synthetic samples via GPT 4.0, class weighting, SMOTE [42] in which synthetic samples are generated for minority classes in the fused feature space via interpolation, and focal loss [175] to handle hard-to-classify instances.

Team AlessioPittiglio [253] explores a wide set of transformer-based text and audio encoders. For text, they evaluate BERT, RoBERTa, DeBERTa and ModernBERT. In particular, the authors propose three different strategies to integrate context information by (i) concatenating tokens during tokenisation, (ii) concatenating pooled embeddings, and (iii) leveraging cross-attention. For audio, they evaluate Wa2Vec 2.0, WavLM and HuBERT. Regarding multimodality, they build an ensemble of the best text (RoBERTa with context information) and audio (HuBERT) models via a weighted average of individual encoder logits. Weights are calibrated via Bayesian optimisation [306].

Team CASS [133] encodes texts with RoBERTa and audio recordings using a BiLSTM with extracted MFCCs features as input [198]. For multimodality, they concatenate the pooled embeddings of RoBERTa for text and Wav2Vec for audio. The concatenated embedding is eventually fed to a logistic regression classifier.

Argumentative Fallacy Detection. Team Ambali_Yashovardhan uses RoBERTa for processing text inputs, and DistilHuBERT for encoding audio recordings. Due to memory constraints, they limit audio sequence length to 320,000 samples (approximately 20 seconds at 16 kHz), truncating longer files. Regarding multimodality, the authors adopt a late fusion strategy where a weighted average of each modality model's logits is computed. The weights are learnt during training. To handle class imbalance, they adopt focal loss [175].

Team EvaAdriana [162] explores five transformer-based models for text modality: BERT, RoBERTa, SBERT, ALBERT, and DeepSeek R1. Up to the last four layers of encoder-based transformers are unfrozen for fine-tuning on the task, while DeepSeek is 4-bit quantized to account for the available computational budget. Regarding audio modality, the authors evaluate two models: a CNN with MFCCs features as inputs and Wav2Vec 2.0. Lastly, they combine the best-performing models for text (RoBERTa) and audio (Wa2Vec 2.0) for multimodality. In particular, they concatenate the output of each modality encoder and feed it to a classification head.

3.6.4 Main Findings

The MM-ArgFallacy2025 shared task revealed key insights about multimodal fallacy analysis through binary detection and multi-class classification tasks, highlighting both capabilities and fundamental limitations in this field.

Textual Dominance — Text-based approaches consistently outperformed audio and multimodal alternatives across all teams in AFC. Team NUST achieved a 0.4856 F1-score for classification, establishing transformer-based models, particularly RoBERTa variants, as the most effective. Team EvaAdriana's comparative evaluation in AFD revealed that fine-tuned transformer models substantially outperformed zero-shot approaches, with their task-specific RoBERTa achieving 0.3393 F1-score compared to DeepSeek-R1's zero-shot performance of 0.1567 on the validation set, underscoring the importance of domain-specific adaptation for specialised argumentation tasks.

Audio Processing Challenges — Audio-only approaches consistently underperformed, with Team AlessioPittiglio's best performance of 0.3559 remaining substantially below text baselines despite employing state-of-the-art models including HuBERT, Wav2Vec2, and Whisper. However, the results achieved by the participants demonstrate progress over previous work in extracting fallacy-relevant acoustic features. Technical constraints forced audio truncation to 15-20 seconds, but more fundamentally, acoustic signatures of fallacious reasoning appear too subtle for current speech processing models to reliably capture.

Limited Multimodal Gains — Multimodal approaches failed to deliver expected performance improvements. Team NUST's late fusion achieved 0.4611, only modestly improving over text baselines while requiring high computational cost. This suggests simple fusion strategies are insufficient to capture complex relationships between semantic content and paralinguistic delivery, with textual information overwhelming rather than complementing audio features.

Effective Strategies — Team NUST's success stemmed from prompt-guided augmentation using GPT-4. Team AlessioPittiglio's context integration (3-4 previous sentences) contributed to strong performance, though benefits were primarily textual, suggesting modality-specific strategies are needed.

Class Imbalance Challenge — Severe class imbalance emerged as the primary technical challenge. Binary detection faced 90.8% vs. 9.2% distribution, while classification presented "double imbalance" with *Appeal to Emotion* comprising 59% of fallacious samples versus <3% for minority classes. Team NUST's synthetic data generation proved more effective than algorithmic adjustments, indicating quality augmentation outperforms technical modifications for addressing imbalance.

3.6.5 Discussion

The MM-ArgFallacy2025 shared task established important baselines while revealing both progress and limitations in multimodal fallacy analysis, providing crucial insights for future research in MAM.

Audio Improvements and Speaker Dependency — While text-based methods achieved the highest performance across both detection and classification tasks, teams demonstrated notable improvements in audio-only approaches compared to our previous work in Section 3.4, with Team AlessioPittiglio achieving a 0.3559 F1-score for classification using HuBERT-based models. These advances suggest that audio modalities contain valuable information for fallacy detection, though current extraction techniques remain limited. The challenges in audio-only approaches may also be attributed to the inherently speaker-dependent nature of acoustic cues. As noted by previous work, different speakers have varying skills in using vocal cues such as articulation, sonority, and tempo, and possess different levels of persuasive power, with vocal characteristics directly affecting the clarity, credibility, and receptivity of a speaker's message [177].

Multimodality Fusion Strategies — The reliance on simple concatenation and late fusion approaches in the proposed multimodal systems reveals fundamental gaps in current methodologies. These approaches fail to capture complex interdependencies between linguistic content and paralinguistic delivery, which likely explains the limited performance of many multimodal systems. Rather than indicating the ineffectiveness of multimodal integration itself, these results highlight the need for more sophisticated fusion architectures that can jointly learn complementary cues from both modalities during training. Future work should prioritise advanced fusion architectures that enable joint learning across modalities, moving beyond late fusion toward cross-attention mechanisms and early integration strategies.

Conditional Audio Generation — Key contributions of this shared task include successful prompt-guided data augmentation and systematic context integration strategies that proved effective for addressing severe class imbalance and improving classification performance. Building on the success of synthetic textual data generation, a possible direction involves controllable conditional audio generation that transcends the basic text-to-speech approaches employed by teams in this shared task. Drawing inspiration from recent advances in signal-to-language aug-

mentation [155] that enable fine-grained control over acoustic parameters such as loudness, pitch, reverb, brightness, and duration, future research could develop fallacy-aware audio generation systems. Such approaches go beyond traditional digital signal processing by incorporating learned representations that capture how acoustic characteristics convey persuasive intent and logical flaws in context. For appeals to emotion, generation could emphasize particular intonation patterns and vocal intensity, while deceptive reasoning patterns could incorporate vocal stress indicators, hesitation markers, or pitch variations suggesting uncertainty.

Limitations and Future Directions — Several key limitations emerged from this shared task, providing valuable guidance for future research. Consistent with our findings in Section 3.4, we argue that fallacy classification should be approached as a genuinely multimodal problem. However, the annotation methodology adopted in this study, following [94], relied exclusively on textual information for both training and evaluation. This text-centric approach overlooks essential acoustic cues in spoken debates that often convey fallacious reasoning. Realising the full potential of multimodal fallacy detection will therefore require the development of new annotation protocols that integrate linguistic and acoustic dimensions from the outset.

The MM-USED-fallacy dataset also faces structural constraints that affect model training and evaluation. With only 3,338 instances for the AFC task, it remains relatively small for deep learning applications, reflecting the broader challenge of obtaining high-quality argumentation annotations. The expansion carried out for this shared task added just 229 additional instances from two 2024 debates, preserving the dataset's limited scale. Moreover, both detection and classification tasks suffer from pronounced class imbalance: fallacious sentences constitute only 9.2% of instances, while within-class distribution is heavily skewed toward Appeal to Emotion, which represents nearly 60% of fallacious cases. Some fallacy types, such as Slippery Slope, are entirely absent from the test set, limiting comprehensive evaluation. These scale and distribution imbalances hinder model generalization across speaking styles, debate contexts, and fallacy types.

MAMKit served as the delivery platform for the shared task datasets and facilitated community experimentation, but it remains under active development. The toolkit currently supports only PyTorch, which may limit adoption by researchers using other frameworks. Its coverage of multimodal resources is also partial. Furthermore, MAMKit presently supports only text and audio modalities, while

the inclusion of visual argumentation resources remains a key future direction. Nonetheless, deploying MAMKit for this shared task enabled the collection of valuable community feedback and highlighted priority areas for improvement.

Overall, the MM-ArgFallacy2025 shared task confirms that fallacy detection and classification remain challenging problems with considerable room for progress. While text-based approaches currently yield the strongest results, multimodal systems require architectural innovation and richer datasets to realise their full potential. Future work should focus on expanding multimodal corpora across election cycles, languages, and speaker demographics, while developing balanced annotation protocols that reflect the diversity of fallacious reasoning. Ultimately, the goal is to build integrated systems that combine semantic and paralinguistic understanding, supporting more transparent analysis of political discourse and contributing to democratic literacy.

3.7 Conclusion

This chapter presents a comprehensive exploration of MAM, tracing a research trajectory from foundational dataset development through specialised fallacy classification to toolkit standardisation and community-wide evaluation. Through four interconnected contributions, we establish the infrastructure and empirical evidence demonstrating the potential of multimodal approaches to advance argumentative analysis in political discourse.

Our research begins with MM-USED, a corpus of 26,781 annotated sentences with corresponding audio samples spanning nearly six decades of US presidential debates. As detailed in Section 3.3, this dataset defines critical benchmarks for core AM tasks — ASD, ACC, and ARC — showing that embedding-based audio encodings consistently outperform feature-based ones and that multimodal integration improves performance across datasets and tasks.

Building on these insights, Section 3.4 introduces MM-USED-fallacy, the first corpus dedicated to detecting and classifying logical fallacies in spoken political discourse. Multimodal integration yields gains of 4-8 F1-score points across models, confirming that argumentative phenomena involving persuasion or emotion are more effectively captured when semantic and paralinguistic cues are combined.

Section 3.5 presents MAMKit, a comprehensive toolkit that standardises access to datasets, models, and evaluation protocols while introducing three fusion strategies: CSA (Concatenation-based Self-Attention), Ensemble (weighted late fusion), and Mul-TA (crossmodal attention). The unified evaluation framework provides new insights into fusion effectiveness and establishes reproducible benchmarks for future research.

Finally, the MM-ArgFallacy2025 shared task validates our findings at a community scale. The competition demonstrates notable advances in audio-based and multimodal approaches, confirms the value of the proposed architectures, and highlights promising directions for deeper multimodal integration.

Together, these contributions underscore the essential role of multimodality in capturing the full complexity of argumentative communication. Yet, progress depends on developing annotation methodologies that integrate acoustic information from the outset. Paralinguistic cues such as intonation, stress, and prosodic emphasis convey layers of argumentative meaning that text alone cannot capture.

Our findings offer practical guidance for future research. Embedding-based

audio representations emerge as the most effective for argument analysis. The resources and frameworks provided through MAMKit and the shared task establish a foundation for systematic, reproducible progress in the field.

Future work should focus on integrating paralinguistic cues into annotation and modelling pipelines, advancing fusion architectures, and exploring synthetic data generation to mitigate data scarcity. Beyond technical innovation, the standardisation and community collaboration initiated here demonstrate that MAM is a mature and growing research area.

Ultimately, MAM opens new avenues for understanding how people argue and persuade in spoken contexts. The datasets, models, and tools developed in this work demonstrate that multimodal approaches can capture the interplay between semantic and paralinguistic signals that text alone overlooks. By establishing shared resources and benchmarks, this research lays the foundation for future systems that enhance the transparency, interpretability, and societal relevance of argumentative analysis.

Link to Research Questions. This chapter primarily addresses **RQ1**, **RQ3**, **RQ4**, and **RQ5**, providing both cross-domain and domain-specific contributions to multimodal argumentative analysis.

- **RQ1: Multimodal Necessity** — Findings validate the necessity of multimodality for analysing persuasive discourse, showing that paralinguistic cues such as prosody, emphasis, and tone capture argumentative phenomena beyond textual content.
- **RQ3: Systematic Solutions** — The development of MAMKit and the MM-ArgFallacy2025 shared task contributes to systematic and reproducible multimodal research, establishing standardised datasets, model architectures, and evaluation frameworks.
- **RQ4: Data Alignment Challenge** — Addressed through the creation of MM-USED and MM-USED-fallacy, which demonstrate how text-based argument mining resources can be effectively aligned with audio content to form comprehensive multimodal political discourse datasets.
- **RQ5: Multimodal Performance** — Empirical evidence across multiple tasks and architectures confirms improvements when integrating acoustic information, proving the advantage of multimodal approaches for argument structure detection and fallacy classification.

CHAPTER 4

Multimodal Clinical Decision Support

4.1 Introduction

The application of artificial intelligence to clinical decision support represents one of the most promising yet challenging frontiers in machine learning. There is a widespread consensus across the research community, funding bodies, and clinical practitioners that addressing complex healthcare challenges requires a multimodal approach. This recognition is reflected in major European research initiatives and funding calls, which increasingly prioritise multimodal AI systems capable of integrating diverse clinical data sources. The imperative for multimodality stems from a fundamental observation: in clinical settings, particularly in diagnostic interviews for mental health and neurological disorders (MHND), human expression is inherently multimodal, spanning speech prosody, linguistic content, and neural activity.

A clinician assessing a patient for depression doesn't rely solely on what the patient says; they attend to how it's said (vocal quality, speech rate, pauses), the content of their words (negative self-talk, cognitive patterns), and observable

behavioural markers [62]. Similarly, when evaluating a patient with Parkinson's disease, neurologists listen for characteristic speech impairments while considering the patient's verbal descriptions of symptoms and, when available, neurophysiological measurements. These clinical scenarios illustrate why AI systems for healthcare must be multimodal by design. Just as no single diagnostic test definitively identifies a complex disorder, no single data modality captures the full picture of a patient's condition. Different modalities provide complementary information: speech prosody may reveal affective states, linguistic content may expose cognitive patterns, and neural signals may indicate underlying neurophysiological changes [252, 283, 33]. Effective clinical AI must integrate these channels just as human clinicians do.

Yet developing trustworthy multimodal systems for healthcare faces three critical, interconnected challenges that distinguish this domain from other AI applications. First, data quality and ethical considerations create unique constraints on dataset creation, sharing, and use. Second, effective multimodal fusion requires principled approaches to combining heterogeneous signals with different temporal dynamics, noise characteristics, and information content. Third, model interpretability becomes non-negotiable in clinical contexts where opaque predictions cannot be deployed, regardless of their accuracy. These challenges are particularly acute in mental health and neurological disorder research, where sensitive data, small sample sizes, vulnerable populations, and the absolute requirement for clinical transparency converge.

4.1.1 Challenges

Data, Ethics, and Quality — Conventional AI-system development has historically followed a model-centric lifecycle, whereby researchers iterate on architectures and training procedures to improve performance while benchmark datasets remain mostly unchanged [380]. However, a recent trend recognizes that data collection and curation is a pivotal yet frequently overlooked step in machine learning pipelines, advocating a fundamental shift from model design to data quality and reliability.

This shift is especially critical in healthcare. Issues originating from problematic data collection practices have been demonstrated across various domains [247, 288, 249]. Non-representative or inaccurate training data can profoundly impact model performance, propagating biases into subsequent development stages and

ultimately leading to negative societal impacts [232, 6, 244]. Furthermore, neglecting to follow adequate policies and protocols for data collection, coupled with these data biases, can end up compromising the system's overall reliability and trustworthiness [343, 173].

The challenges are especially pronounced in domains targeting MHND.[1] Data may be used to develop AI systems that identify high-risk populations for early intervention, detect and predict disorders in decision support systems, and trigger targeted actions, potentially improving access to mental health care [82]. For instance, digitally-delivered targeted interventions could mitigate the burden of mental illness in hard-to-reach populations. However, creating datasets for these applications may be potentially harmful for all stakeholders. Individuals could volunteer sensitive content when speaking spontaneously during data collection. Such risks must be accounted for and dealt with systematically, for example by obtaining proper participant consent, ensuring data security and confidentiality, and mitigating biases [84, 203, 82]. Moreover, to enable broader accessibility of the datasets, a delicate balance must be maintained between privacy protection and transparent data utilization [244].

Recognising the growing concern surrounding these sensitive topics, previous research has addressed ethical issues at different stages of the development process [5, 95, 327]. However, there is a noticeable gap in the literature concerning speech datasets for MHND.

To address this, the first contribution of this chapter (Section 4.3) defines a set of desiderata for ethically responsible speech data collection and curation in mental health and neurological research. Building on existing guidelines and recent analyses of ethical concerns, we distil these principles into an actionable checklist to support researchers during dataset design, implementation, and maintenance. We then apply this framework to critically review 36 speech-based datasets, identifying recurring pitfalls and highlighting areas for improvement. Although centred on MHND, the proposed framework generalises to other domains involving speech-based data.

Multimodal Clinical Data Integration — Given responsibly collected data, the next challenge lies in building systems that effectively leverage multiple modalities. The motivation is clear from both clinical practice and the nature of MHND them-

[1] In the clinical literature, a clear distinction between symptoms and disorders is often delineated, yet the majority of datasets in the domain of MHND have primarily focused on detecting disorders, albeit with some contributions towards symptom detection, such as aphasia [185].

selves. Depression, for instance, is a widespread mental health condition predicted to become the second leading cause of disease burden by 2030 [204], with COVID-19 causing a 27.6% rise in global cases [287]. Similarly, Parkinson's disease is a progressive neurodegenerative disorder primarily marked by the deterioration of dopaminergic neurons in the midbrain, leading to motor and non-motor symptoms including tremors, bradykinesia, cognitive impairment, and depression [116, 254]. Importantly, during the prodromal stages of Parkinson's disease, patients often exhibit speech impairments that can serve as early indicators [252, 283]. Given the non-invasive, cost-effective, and automated nature of speech analysis, researchers have increasingly focused on developing automated detection systems to support clinical decision-making, enable telemedicine applications, and facilitate early detection [157].

Multimodal approaches have gained particular attention, motivated by the fact that in clinical settings such as diagnostic interviews, human expression is inherently multimodal, spanning speech, language, and neural activity. However, current studies often suffer from critical methodological gaps, including limited modality integration, inconsistent evaluation protocols, and potential data leakage, which hinder reproducibility and the fair assessment of model performance.

Existing work reveals that models leveraging two modalities dominate the field, with approaches exploring combinations such as EEG and speech, speech and text, or EEG and facial expressions. State-of-the-art performance in multimodal depression detection, for instance, ranges from 85% to 97%, depending on the dataset and modality combinations. However, these bimodal approaches constrain their potential by overlooking comprehensive trimodal integration. Moreover, most studies exclude one or more key modalities and lack transparent data-splitting protocols that would ensure proper subject-level validation. True trimodal approaches, which simultaneously integrate neural, acoustic, and linguistic information, remain scarce, and existing attempts have not systematically compared handcrafted features against pretrained embeddings (including brain-pretrained models) or explored diverse fusion strategies with rigorous cross-validation protocols.

The second contribution of this chapter (Section 4.4) addresses these limitations by introducing a comprehensive framework for multimodal depression detection that integrates EEG, speech, and text. Using consistent subject-level validation, we compare feature representations, deep learning architectures, and

fusion strategies to identify which modalities contribute most to diagnostic performance and robustness. This case study demonstrates the diagnostic potential of trimodal fusion while establishing reproducible benchmarks for future research in clinical multimodal AI.

Explainability and Clinical Trust — Despite substantial advances in classification accuracy using multimodal approaches, model interpretability remains critically underexplored, particularly for speech-based detection. In clinical settings, XAI is essential for providing clear, clinically relevant insights that are crucial for the acceptance of automated systems in clinical trials and eventual deployment in healthcare practice.

The challenge is particularly acute for speech and audio modalities. While text-based models can provide interpretable explanations by highlighting important words or phrases, speech presents a fundamentally different problem. Standard representations (spectrograms, mel-frequency features) are complex, non-intuitive visualisations that convey little semantic meaning to domain experts. A clinician examining highlighted regions of a spectrogram gains no intuitive understanding of what acoustic properties the model considers diagnostically relevant.

While many post-hoc explanation methods exist — including perturbation-based approaches that assess feature importance and gradient-based methods that compute input attribution [218, 205] — a critical question remains: do these model-agnostic techniques, when applied to speech classifiers, provide explanations that are truly informative for clinical decision-making, or do they merely achieve technical faithfulness to the model's internal representations without conveying actionable insights to domain experts?

The third contribution of this chapter (Section 4.5) investigates this interpretability gap by evaluating prominent explainability techniques in the context of Parkinson's disease detection from speech. Through systematic comparison, we assess whether existing methods provide explanations that clinicians can meaningfully interpret. The findings reveal a consistent disconnect: current techniques achieve high technical faithfulness but limited clinical relevance. This underscores the urgent need for perceptually grounded approaches that align model reasoning with human auditory understanding, an essential step toward building trustworthy multimodal clinical AI.

4.1.2 Our Contribution

This chapter presents three interconnected contributions that address the aforementioned challenges and advance the state-of-the-art in multimodal clinical decision support:

- We establish ethical guidelines and best practices for the creation and curation of speech datasets in mental health and neurological disorder research. Recognising that data quality fundamentally shapes model trustworthiness in clinical applications, we define a comprehensive set of desiderata for responsible speech data collection from vulnerable populations. We operationalise these principles through an actionable checklist and apply it to critically examine 36 published speech datasets across multiple mental health and neurological conditions. This work addresses the critical gap in ethical frameworks specifically tailored for clinical speech data, providing both an analytical tool for informed use of existing resources and a practical guide for future dataset construction that prioritises patient privacy, informed consent, and equitable representation.

- We advance multimodal depression detection by conducting the first systematic exploration of trimodal integration across EEG, speech, and text modalities. Our work addresses three fundamental limitations in existing research: the predominance of bimodal approaches that overlook comprehensive multimodal integration, the lack of systematic comparisons between handcrafted features and modern pretrained embeddings (particularly brain-pretrained models for EEG), and inconsistent evaluation protocols that compromise generalisation assessment. We establish rigorous experimental protocols using stratified subject-level cross-validation and demonstrate that carefully designed trimodal models outperform both unimodal and bimodal configurations, achieving state-of-the-art results. Beyond identifying optimal architectures, we introduce a reproducible experimental framework that serves as a reference setup for future research.

- We systematically evaluate explainability methods for speech-based clinical detection, using Parkinson's disease as a case study. Through comprehensive assessment of perturbation-based and gradient-based explanation methods across multiple evaluation dimensions, we reveal a critical disconnect: while explanations achieve high technical faithfulness to classifier decisions,

they often fail to provide insights that healthcare professionals can readily interpret. This finding has profound implications for clinical deployment, demonstrating that technical faithfulness alone is insufficient and that interpretability for domain experts represents a fundamental unsolved challenge in speech-based clinical AI.

4.1.3 Chapter Structure

This chapter addresses the aforementioned challenges in sequence, recognising that each builds on the previous. We begin by establishing ethical and methodological foundations for data curation and collection (Section 4.3), proceed to demonstrate effective multimodal fusion strategies (Section 4.4), and conclude by revealing critical gaps in current explainability approaches (Section 4.5). Together, these contributions trace the full pipeline of responsible clinical AI development: from ethical data foundations, through effective multimodal modeling, to interpretable explanations. While each addresses specific technical questions, they collectively demonstrate that technical performance must be balanced with ethical considerations and clinical utility. The insights from this chapter apply not only to MHND but to any clinical application of multimodal AI. The path toward trustworthy clinical decision support systems requires addressing all three challenges in concert. High-performing models built on ethically compromised data cannot be deployed responsibly. Perfectly interpretable models that perform poorly are clinically useless. And accurate, well-explained models trained on biased data may systematically disadvantage vulnerable populations. Only by simultaneously attending to data quality, model performance, and interpretability can we develop AI systems worthy of clinical trust and capable of meaningful impact on patient care.

4.2 Related Work

This section reviews existing work relevant to the three main challenges addressed in this chapter: ethical data collection and documentation, multimodal integration for clinical detection, and explainability methods for speech-based systems.

4.2.1 Ethical Concerns and Dataset Documentation

Ethical Concerns in Mental Health and Neurological Disorder Research — To the best of our knowledge, there are no desiderata covering the creation and collection of speech datasets in MHND research. However, a rich literature offers analyses of ethical concerns and best practices for dataset construction in related domains.

Working at the intersection of social media and mental health, [84] examines the ethical issues arising from the collection, storage, and analysis of georeferenced tweets. They identify four issues and suggested methods to address them: (1) *users' privacy expectations and informed consent* - stressing the importance of informed consent when collecting data as well as ensuring that participants understand the implications of sharing data in a public online domain; (2) *data security* - focusing on technological solutions like encryption; (3) *data confidentiality* - related to (1), this point also focuses on the importance of informing participants how their data is stored and transferred, as well as the potential risk of breaching the data confidentiality; and (4) *social responsibility* - advocating for special consideration and respect towards social groups and structures that may be disrupted by the proposed research (e.g. stigmatizing users by identifying them as diagnosed with a mental condition).

[203] extends the study's scope from social media to a broader set of phenotyping tools for mental health applications, including biometric and personal data from smartphones and wearables. In their analysis of ethical concerns in this field, they relied on a panel of experts in digital phenotyping, data science, mental health, law, and ethics. The following concerns emerged: (1) *privacy and data protection*; (2) *transparency* of the research processes, risks, limitations, and results; (3) informed *consent* of everyone whose personal data is collected; (4) *oversight and accountability* by way of external ethical reviews; (5) *fairness and bias*; and (6) *validation of tools* by establishing the evidence of validity for the use of said tools.

A broader analysis of the ethics and law in technological research in mental health care is provided by [95]. They surveyed empirical scholarly literature on the application of algorithmic and data-driven technologies (including social media, smartphones, sensing technology, and chatbots) in mental health initiatives to identify the legal and ethical issues that have been raised. Most of the literature discussed privacy, mainly in terms of respecting the privacy of research participants. A small number of studies discussed ethics directly and indirectly, whereas legal issues were not substantively discussed in any studies, although some legal issues were discussed in passing, such as the rights of user subjects and privacy law compliance. From the papers that discussed ethics, the authors summarised the following ethical concepts: (1) *privacy*; (2) *security*; (3) *safety*; (4) *transparency*; (5) *autonomy*; and (6) *justice*.

On the topic of AI in mental health applications, several reviews emphasise critical ethical concerns that must be addressed [280, 327, 130, 102, 382, 328]. Key issues include *accountability* and *transparency*, where AI systems must offer interpretable decision-making processes to ensure trust among patients and clinicians; *privacy* and *data security*, highlighting the need for strong protections against unauthorized access to sensitive mental health data; *bias* and *equity*, as AI may perpetuate societal inequalities if trained on biased data, leading to unequal care; and *autonomy* and *informed consent*, ensuring that AI respects patient autonomy while balancing their needs with available resources. Additionally, concerns about the lack of empathy, human connection, and continuity of care in AI-driven systems stress the importance of human oversight to ensure AI supports, rather than replaces, human professionals.

Recommendation Checklists — Several contributions have highlighted that many apparently positive findings in machine learning applications were actually caused by methodological errors [134] that may occur in the experimental setup and in data collection [247, 288, 249]. Moreover, it is well-known that faulty data collection strategies may lead to biased machine learning models, whose application in real-life scenarios may have harmful implications on people (e.g., discrimination and stereotypes) [27, 28].

Besides the risks these factors pose to a machine learning model and, indirectly, to its consumers, researchers have highlighted the importance of thoroughly documenting methodology to foster research reproducibility, awareness of technology consumption, and awareness of potential ethical issues. In par-

ticular, literature work has proposed recommendations and desiderata targeting these factors through standardised sheets, checklists, and '*how-to*' cards. Notable examples include datasheets for documenting and tracking data collection [90, 256, 316, 383] and its derived software [20], cards and factsheets for model selection [215, 1, 307, 10], ethics sheets for evaluating AI tasks [217] and their social impact [297, 68], sheets for instructing on human evaluation [300] and annotation [71], and checklists for paper submission [190].

These tools aim to instil technology awareness in daily machine learning system practitioners and promote reproducibility. However, they are often meant as starting points for critical thinking; therefore, they are intentionally broad and general-purpose and may not capture factors tailored to specific domains. For this reason, ad-hoc extensions of such recommendations have been proposed targeting healthcare datasets [279] and ethical and data collection issues on speech-language technology [244, 333, 88, 191]. In the domain of speech datasets concerning language impairment, [359] address aphasia providing a short list of requirements.[2]

To the best of our knowledge, we are the first to specifically address the domain of speech datasets for MHND.

4.2.2 Multimodal Depression Detection from Clinical Interviews

The related work reviewed here focuses specifically on multimodal approaches to depression detection using data from clinical interviews, where multiple modalities (EEG, speech, text) are recorded during structured diagnostic sessions.

Bimodal Approaches — Most existing studies employ two modalities. The MODMA dataset [36], which contains EEG recordings and speech audio from clinical depression assessments, has become a primary benchmark for evaluating EEG–speech integration methods. For instance, Yousufi et al. [375] applied DenseNet121 to EEG and speech spectrograms from MODMA, while Qayyum et al. [258] employed Vision Transformers on EEG–speech data. Other bimodal studies explored EEG–speech fusion using graph convolutional networks on MODMA [124], and speech–text integration on the E-DAIC dataset [97], which contains audio and transcribed interviews collected in clinical settings, using CNN-Long Short-Term Memory (LSTM) attention mechanisms [234]. Tiwary et al. [329] investigated EEG–facial expression fusion through a late-fusion architecture trained on dis-

[2] Aphasia can be a symptom of a neurological disorder or brain injury, but is not considered a disorder in itself. For this reason, we do not address it in our work.

tinct datasets: AVEC 2018 [277] for visual modelling, MODMA for EEG-based LSTM training, and DEAP [148] for cross-dataset evaluation combining EEG and video modalities. In [64], an extensive comparison of speech–text fusion techniques was conducted on E-DAIC, although EEG was not included.

Overall, state-of-the-art performance in multimodal depression detection from clinical interviews ranges from 85−97%, depending on the dataset and modality combinations.

All the aforementioned approaches comprise only two modalities, thereby constraining their potential by overlooking trimodal approaches. Moreover, most of them exclude text modality and lack transparent data-splitting protocols.

Trimodal Approaches — Trimodal depression detection has been explored using different modality combinations. Some studies integrate speech, video, and text modalities [273], employing multi-level attention mechanisms to fuse features from these three sources. However, such visual-based approaches are outside the scope of the present work, as we do not incorporate visual modalities. To the best of our knowledge, only a few studies have investigated trimodal approaches involving EEG, and one representative example is that of He et al. [109], who integrated speech, EEG, and text using GAT-CNN-MpNet architectures on the MODMA dataset. Their model achieved approximately 90% balanced accuracy through weighted late fusion, although the study did not compare handcrafted and pre-trained features, and explored only basic fusion strategies. Moreover, the paper did not specify whether 5-fold cross-validation was conducted at the segment or subject level, a critical distinction that determines whether results reflect genuine generalisation to new patients or merely memorisation of training subjects.

4.2.3 Explainability in Speech-Based Clinical Detection

To investigate the effectiveness of explainability methods for speech-based clinical detection, we use Parkinson's disease as a case study. At the time this research was conducted, Parkinson's detection offered an ideal experimental setting with publicly available datasets and established baselines suitable for evaluating interpretability. The subsequent depression detection work presented in Section 4.4 builds on these methodological insights. Importantly, the approaches developed here generalise beyond Parkinson's disease to other clinical domains involving speech, such as depression, Alzheimer's disease, and broader mental health

and neurological conditions. They also extend to non-clinical applications—including emotion recognition, speaker identification, and speech enhancement—where model interpretability remains crucial.

The wealth of research on automated speech analysis in Parkinson's disease has established it as an ideal testbed for interpretability research. Acoustic impairments can serve as reliable biomarkers of disease progression, and models based on convolutional neural networks (CNNs) trained on spectrograms have been successfully used to detect dysarthria, assess severity, and distinguish between patients and healthy controls [308, 278, 344, 261]. However, while deep models using non-interpretable features often achieve higher accuracy, they offer limited transparency for clinical interpretation [87].

Although this limitation has spurred growing interest in XAI, most work in speech-based clinical detection continues to prioritise performance over interpretability.

Existing approaches typically employ post-hoc visualisation techniques, such as GradCAM [293], EigenCAM [222], or SHAP [188], but these methods often lack rigorous quantitative evaluation and rely on low-level acoustic features, like mel-frequency coefficients, that do not align with perceptually meaningful clinical concepts.

Recent research in audio interpretability has proposed methods to address this gap. Techniques such as layer-wise relevance propagation [19], Guided Backpropagation [220], and additive-noise perturbation [335, 136] have improved the understanding of model attention within spectrograms. More recent work, including SLIME [214, 213], AudioLIME [108, 54], Listen-to-Interpret (L2I) [245], and L-MAC [242], has introduced perceptually oriented frameworks that generate explanations in both spectrogram and time domains, emphasising the importance of human-auditory alignment.

Despite these methodological advances, a fundamental question remains unanswered: do existing post-hoc explainability methods — whether traditional or perceptually-oriented — provide insights that are genuinely useful for clinical decision-making? Our work (Section 4.5) addresses this gap by systematically evaluating representative post-hoc techniques in Parkinson's disease detection, revealing a disconnect between technical faithfulness and clinical interpretability that motivates the need for perceptually grounded approaches.

4.3 Promoting the Responsible Development of Speech Datasets for Mental Health and Neurological Disorders Research

This section presents our work on establishing ethical guidelines and best practices for speech dataset creation in mental health and neurological disorder research, conducted in collaboration with Ana Tanevska, Andrea Galassi, Alessio Galatolo, Federico Ruggeri, and Paolo Torroni [201]. As outlined in the chapter introduction (Section 4.1), the shift toward data-centric AI has revealed that issues in data collection practices can profoundly impact model performance and propagate biases, ultimately compromising system reliability and trustworthiness — challenges that become particularly acute in sensitive clinical domains.

Key Contribution: Despite growing awareness of these concerns, a significant gap persists in the literature regarding ethical frameworks specifically tailored for speech datasets in mental health applications. Our contribution addresses this gap by defining a set of desiderata — desired features and attributes to guide responsible speech data collection efforts. These desiderata serve dual purposes: as an analytical tool for researchers committed to more informed use of existing resources, and as a practical guide for future dataset construction. We operationalise these principles through an actionable checklist and apply it to critically examine 36 published speech datasets across multiple mental health and neurological conditions.[a]

[a] Additional survey materials and datasets details are available at `https://github.com/nlp-unibo/ethical-survey-speech`

To ground this work in concrete realities rather than abstract principles, we begin with a detailed case study of the DAIC-WoZ [97] depression dataset. This choice is strategic: DAIC-WoZ represents one of the oldest and arguably most influential datasets in the field, having played a pivotal role in advancing depression detection research. Its widespread adoption over many years has surfaced various challenges — from demographic imbalances and unclear labelling criteria to maintainability issues — that exemplify the kinds of problems our desiderata aim to prevent. By examining DAIC-WoZ, we illustrate how even well-intentioned datasets can develop limitations that affect downstream research, and we demon-

strate the value of systematic ethical reflection throughout the dataset lifecycle. Importantly, our goal is not to critique DAIC-WoZ specifically, but rather to use it as a lens through which to identify patterns that may exist across the field and to motivate the comprehensive framework that follows.

4.3.1 A Case Study on DAIC-WoZ Depression Research

DAIC-WoZ is a subset of the larger Distress Analysis Interview Corpus (DAIC) [97] where a person is participating in a semi-structured interview to assess their risk for Post-Traumatic Stress Disorder (PTSD) and Major Depressive Disorder (MDD). The datasets consist of the data from the interviews, such as the speech data and the recordings' transcriptions, and is annotated with the Patient Health Questionnaire (PHQ) for depression [152]. The dataset has been vastly used for the training of systems aimed at automatic depression detection [13, 272, 367, 80, 127]. In these works, authors develop methods to automatically recognise depression (primarily as a binary classification task) from the audio of the interviews.

DAIC-WoZ presents an imbalance both in terms of class distribution[3] (65% negative, 35% positive) as well as gender/sex distribution (54% females, 46% males). Bias and lack of diversity in a dataset may propagate to the models using them, often compromising their generalization abilities. [13] analyses the effect of gender/sex bias on model performance and shows how some features like mel-spectrogram, which are commonly used in several works, are not robust to it. Although it may be possible to partially address these issues through ad-hoc methods [13, 193], a more effective approach would be to adopt a rigorous and well-balanced data collection process from the start.

Another issue concerns the meaning of annotations. DAIC-WoZ includes two labels: the PHQ-8 score [152] and a corresponding binary label for depression. In the paper presenting DAIC-WoZ, [97] do not specify the criterion used to map the PHQ-8 score into the binary label.[4] While most research involving this dataset has followed [153] and used the label "depressed" for instances with scores > 9, [367] categorize as "depressed" instances with a score > 10, obtaining results that are not directly comparable with previous literature. Clarity in the definition and

[3] In this work, we use the term *class* in the sense of category, as it is common in machine learning when addressing classification tasks.

[4] Nonetheless, in the documentation of the dataset, they specify they assigned the label "depressed" to instances with scores > 9.

motivation of the labels used in the dataset would favour future researchers in coherently structuring the task and may prevent inconsistencies in the experimental results. This would especially favour researchers who do not have a medical background.

A final issue worth mentioning is maintainability. Often, datasets have known issues with some instances. DAIC-WoZ is no exception. For example, [13] identify labeling errors, out-of-sync segments, and missing transcriptions.[5] Errors are difficult to avoid and may only become evident after the dataset has been made public. However, a solid distribution infrastructure supporting maintainability may help address this problem effectively. To the best of our knowledge, there is no method to report these issues or update the dataset with corrections, which inevitably puts unaware researchers and general users of the dataset at a disadvantage.

These limitations of DAIC-WoZ, which have emerged over many years of use of the dataset, outline examples of problems that may also be present in other datasets. Therefore, the definition of best practices and desiderata for speech datasets in the MHND domain could induce a principled approach to creating new resources and a more aware use of existing ones.

4.3.2 Desiderata

Our analysis leads us to define a of list desiderata composed of: (i) informed consent, (ii) data storage and security, (iii) data privacy, (iv) accountability, (v) fairness, bias, and diversity, (vi) data quality, validation, and maintainability, (vii) discourse genre. Since the legal framework of reference will depend on the country where the data is collected, our list does not include normative obligations, which nevertheless should be properly taken into account by researchers.

We shall remark that although we treat desiderata as separate concepts, there is often some overlap in the literature. For example, [84] discuss the topic of informing participants both about *users' privacy expectations and informed consent* and *data confidentiality*.

Informed Consent

In the context of clinical and health research environments, it is crucial for consent procedures to thoroughly communicate to individuals the specifics of when and how their data is being collected and utilized [203].

[5] `https://github.com/adbailey1/daic_woz_process`

In this domain, informed consent can be defined with the following three key components - information, comprehension, and voluntariness [5]. Participants, especially in sensitive areas like MHND, may not anticipate their speech being observed and analyzed for research purposes. Thus, securing informed consent ensures their understanding and agreement with the research parameters. This commitment extends beyond the initial consent to subsequent phases, where data sharing and reuse may occur, underlining the ethical responsibility to respect participants' autonomy.

Moreover, the collection, utilization, and sharing of speech data, particularly identifiable information, are contingent upon explicit and ongoing informed consent. Participants should retain the right to withdraw consent at any point, acknowledging the sensitive nature of MHND-related data.

Transparency is pivotal in this domain. Comprehensive information concerning all the processes in which data will be involved, such as collection, storage, and analysis, should be communicated transparently. This includes the genres of speech data collected, potential inferences, intended usage, generated reports, recipients of data and reports, associated risks, benefits, and limitations.

Stakeholders, including participants, clinicians, and researchers, should have accessible explanations tailored to their specific interests and concerns. This comprehensive approach to consent and transparency fosters an ethical foundation for speech dataset utilization in MHND research.

Data Storage and Security

In previous research on mental health, the importance of data storage and security in digital contexts has been underscored. Recommendations include employing encryption and involving IT security teams to protect online activities [84]. Moreover, transparent communication about data handling practices, especially regarding high-risk scenarios, is essential [84]. Clear security standards and regular audits are crucial to monitor compliance and address vulnerabilities [203]. Additionally, the transition to digital platforms poses increased risks of data leakage and loss, raising concerns about privacy and trust in mental healthcare [280].

In light of these considerations, for MHND speech datasets, we advocate for stringent security measures. This includes ensuring secure storage through encryption and also addressing the scenario where data, while not openly accessible

through open access, may still be retained and accessible upon request.

In instances where data are transferred through different devices, such as the collection of speech datasets through crowdsourcing, the focus on data security also extends to these transmissions. It is crucial to explicitly outline and adhere to secure transport encryption protocols, reinforcing a comprehensive approach to safeguarding sensitive data throughout the entire data collection process.

Data Privacy

Existing literature emphasises the need for transparent regulations to address privacy concerns [364], with granular consent proposed as an effective means to empower users in digital settings [140]. Privacy, often intertwined with consent, emerges as a critical aspect in safeguarding users' data [280]. The evolving landscape, encompassing digital phenotyping and various technologies, raises concerns regarding data protection, reidentification risks, and the sale of health inferences [203]. To mitigate these challenges, standards like differential privacy measures are recommended [203]. Notably, the intersection of social media, mobile apps, sensing technology, and chatbots in mental health applications reveals a focus on detection but a limited discussion on ethical implications [95]. When utilising public online data, ensuring informed consent and respecting users' privacy expectations becomes imperative [84].

Our approach involves a deliberate separation of the concepts of privacy and consent. When referring to privacy in our context, we encompass a spectrum of considerations essential for ensuring protection, especially when data are made open access or are accessible to third parties. This involves meticulous attention to factors such as anonymisation and the management of non-identifiable aspects. This commitment to safeguarding privacy extends beyond the initial data collection phase, as it remains crucial when sharing data with collaborating partners. Notably, the responsibility for maintaining privacy measures should be ingrained in every step of the data lifecycle, from the physical collection of data to its management and eventual analysis. This holistic approach underscores our commitment to upholding privacy standards at every stage of the data handling process.

In light of the above, our desiderata include (i) careful avoidance of identifying participants as members of vulnerable groups, (ii) rigorous redaction of privacy-sensitive and identifying information, and (iii) partial or full anonymisation when

sharing data with third parties.

These recommendations aim to foster responsible data practices, acknowledging the evolving landscape of MHND technologies while prioritising user privacy and informed consent.

Accountability

Accountability in the context of AI applications for mental health emerges as a multifaceted concept, with its roots traced through various dimensions in existing literature.

First-wave accountability, observed in the development of mental health apps, concentrates on linguistic inclusivity. This involves ensuring that linguistic corpora of stimuli and responses adequately represent diverse communities with distinct ethnicities and modes of self-presentation [95]. Transitioning beyond linguistic considerations, second-wave concerns adopt a law and political economy approach. This perspective questions whether mental health apps might prematurely disrupt markets and potentially substitute expert professionals with more cost-effective, albeit limited, software. The second wave prompts critical inquiries into the beneficiaries and burdens associated with data collection, analysis, and use [95].

In the development of digital phenotyping tools, ethical considerations stress the importance of an independent interdisciplinary review for potential ethical issues. This review process begins early in the development stages, ensuring responsible and ethical practices [203]. Moreover, in this regard, the imperative to adhere to current and future laws is a central ethical consideration [5].

Insights from discussions on the role of AI in psychiatry further underscore the significance of accountability. While AI holds potential in assisting data analysis and predictions, it is recognised that it cannot fully replace the indispensable human elements of psychiatry, such as empathy and the therapeutic relationship between patients and psychiatrists [327]. Responsible AI use in psychiatry requires a nuanced understanding of its limitations and ethical implications, emphasising transparency and accountability [327]. Against this backdrop, our desiderata for constructing speech datasets for MHND are: (i) conduct thorough ethical review and approval prior to data collection to align with established ethical standards; (ii) collect information on participants' country of residence to ensure dataset compliance with current and future laws, respecting regional ethical guidelines.

Fairness, Bias, and Diversity

Within the broader health landscape, the imperative to address bias, ensure fairness, and promote diversity pervades discussions that extend into the development of speech datasets for mental health [203]. The foundational principles of identifying and minimizing bias in digital phenotyping tools resonate deeply. Collaborative efforts and partnerships are emphasized to navigate potential biases in diverse communities and local contexts, highlighting the need for a collective commitment to fair representation.

Extending beyond digital phenotyping, the discourse on bias mitigation in algorithm development is paramount. This multifaceted approach involves scrutinizing training data, refining algorithmic processes, addressing biases in tool transfer across contexts, and ensuring an unbiased interpretation of digital phenotyping findings. Periodic reviews and re-evaluations are advocated, acknowledging the dynamic nature of bias management.

Diversity in research participants is underscored as a crucial component of equity and fairness. The lack of representation based on race, gender/sex, or disability in research data raises concerns about potential disparities in the development and implementation of digital phenotyping tools. The call for practices that account for the diversity of communities and contexts during design and development processes echoes the importance of mitigating harm to marginalised groups. It is evident from existing literature, particularly in computer science and MHND research, that the distinction between *sex* and *gender* is often conflated, despite their fundamental differences. In these studies, the term *gender* is frequently used to encompass both biological sex and socially constructed roles. [135] are some of the latest from many authors who clarify this distinction, defining *sex* primarily as biological factors and *gender* as socially constructed roles.

It is worth noting, researchers' opinions about this terminology are complex and opposing, with some authors affirming that *sex* should be used when reporting biological differences between males and females, and *gender* for sociocultural differences between women and men [221, 135], but others arguing that the biological and sociocultural factors are closely intertwined, such that sex cannot be studied without considering gender (and vice versa), and thus the distinction between the terms sex and gender should be abandoned [373, 341]. Following the suggestions of [118], in our work, we adopt the term *gender/sex*, both as the simplest way to discuss research which uses the two terms interchangeably, as well as

a recognition of the complexity of the issue. We would, however, like to use this opportunity to reach out to the community and advocate for a more responsible use of the appropriate labels in this area in the future.

As ethical considerations unfold in constructing speech datasets, acknowledging representational and historical biases becomes indispensable. Diversity in speech datasets is multifaceted, encompassing (a) the distribution of languages, (b) varieties within a single language, and (c) speaker socioeconomic and demographic factors [244]. This diversity also involves various socioeconomic and demographic factors like nationality, gender/sex, age, education, income, and even health status. For instance, non-binary individuals and certain age groups, like young children, are often underrepresented [244]. It is important to recognize that while no dataset can fully capture all variations of human speech, it must reflect the specific domain of the intended application. For example, if the target application serves multiple ethnicities, the dataset must represent all these ethnic groups to ensure fair performance across populations. Recognising that biases learned by machine learning models can perpetuate harmful stereotypes emphasises the need for meticulous curation of datasets to avoid reinforcing societal prejudices [5]. The nuances of gender/sex, race, and age in dataset creation reveal instances of disparate algorithmic performance and underscore the imperative to address biases within these dimensions.

Furthermore, the digital divide and accessibility issues add another layer to the ethical discourse. Ensuring inclusivity beyond traditional attributes like age, gender/sex, and race, by considering other sensitive attributes and addressing accessibility concerns, emerges as a crucial facet. To counter confused taxonomies and enhance accuracy, self-reported annotations are proposed, empowering individuals to define their identity and minimizing misclassification. Additionally, providing fair treatment and compensation to contributors emphasizes the ethical responsibility towards those involved in dataset creation.

In the broader context of AI and Big Data in mental health, the implications of bias go beyond technical challenges, extending to potential discrimination and exclusion. The focus on social, gender/sex, and ethnic determinants in training data is crucial for avoiding biased AI systems and fostering a more personalised treatment approach [280].

These considerations converge into our desiderata: (i) ensure fair representation by developing effective strategies for identifying and minimising bias in the

development of speech datasets, ensuring fair representation, such as balancing the distribution of demographics of the participants; (ii) identify and highlight limitations and biases that may be present in the data, discussing potential repercussions; (iii) implement self-reported annotations for diverse identity attributes, provide open-ended response options, acknowledge the multiplicity of identity, and collect aggregate data for commonly ignored sensitive attributes to enhance fairness and inclusivity in speech datasets for MHND; (iv) when conducting research, implement strategies to mitigate bias at different levels of algorithm development, spanning training data, algorithmic processes, tool transfer, and the interpretation of findings.

Data Quality, Validation, and Maintainability

In the realm of AI applications for MHND, the literature often lacks specific discussions on ethical aspects related to data quality, validation, and maintainability. In particular, there was consensus regarding the need to have a mechanism for reviewing or auditing the validity of digital phenotyping tools beyond their initial deployment, such as evaluating software updates or device uses deployed in new contexts [203]. However, while these standards primarily address the application viewpoint, there is a significant gap in considering the connected data. Recognising this gap, we propose guidelines tailored for the construction of MHND speech datasets, synthesising insights gleaned from an extensive literature review. First and foremost, we advocate for a meticulous verification of the speech data's quality to mitigate errors and inaccuracies. This involves ensuring that the dataset faithfully represents the relevant distribution of the intended application domain. In particular, a dataset should reflect the diversity of speech patterns, accents, speech disorders, and demographic factors relevant to the research objectives. This includes capturing a broad range of speech characteristics, such as those seen in individuals with mental health or neurological disorders, and accounting for variations in gender/sex, age, ethnicity, and socioeconomic background. By including such diversity, we reduce the risk of bias in speech recognition models, ensuring they perform more accurately across all demographic and linguistic groups. As discussed in Section 4.3.2, it is important to recognise that while no dataset can fully capture all variations of human speech, it must reflect the specific domain of the intended application.

Moreover, it is essential to describe the recording setting in detail, encompass-

ing environmental specifics, technical aspects of recording instruments (such as sample rate and the presence of noise reduction systems), and the use of personal or environmental microphones. These technical parameters significantly influence data quality and validity, making detailed documentation imperative for researchers and clinicians to assess reliability and relevance.

Annotation of data, especially in the context of MHND, is underscored as a key step. In addition to standard annotation practices (e.g., the usage of self-reported questionnaires to retrieve PHQ-9 scores for depression), involving domain experts with a broad understanding of MHND nuances is crucial for contextualizing and interpreting the data accurately. Recent work suggests that when using self-report measures, the goal of the study must be reframed from predicting diagnosis to predicting self-report questionnaire scores, which may not always match clinical diagnosis [185]. Moreover, specifying the diagnostic manual (i.e., diagnostic criteria) used for reference, such as the DSM-IV or DSM-V for the case of depression, is important for at least two reasons: *clinical interpretation* and *research reproducibility*. Indeed, specifying the diagnostic criteria allows for better clinical interpretation of the findings. Clinicians need to know exactly what criteria were applied to understand the results in the context of their practice. Moreover, explicitly stating the diagnostic criteria ensures that other researchers can reproduce the study accurately. This is important for building upon existing research and validating findings in different populations or settings. For the sake of clarity and completeness, authors should always explicitly specify the diagnostic criteria, even when they could be inferred from other information (e.g., in the case of the PHQ-9 aligning with DSM-IV criteria).

Furthermore, maintainability stands as a pivotal factor in ensuring the reliability of data. In the realm of MHND speech datasets, the acknowledgement that modifications may occur underscores the need for a systematic approach to tracking these changes. This is essential for users who rely on clear and transparent references to specific dataset versions.

A resilient platform for dataset construction plays a central role in facilitating maintainability. This platform should not only support effective versioning, allowing authors to document and manage changes seamlessly, but also incorporate mechanisms for users to report issues. Accessibility is paramount, and a user-friendly platform significantly contributes to ensuring that users can easily locate and utilise the dataset.

Emphasising accessibility goes hand in hand with fostering an environment where users can actively contribute to maintaining and improving the dataset. Therefore, a transparent approach to data maintenance involves not only making the dataset easily accessible but also establishing clear protocols for users to report issues and actively participate in ongoing improvements.

In summary, incorporating versioning and issue-tracking capabilities within the dataset platform is crucial for ensuring continuous usability and enhancement. This not only aligns with ethical considerations but also fortifies the overall robustness and reliability of MHND speech datasets in the ever-evolving landscape of AI applications.

Discourse Genre

In the development of speech datasets for MHND, an additional ethical consideration involves integrating the concept of *discourse genre*. This becomes especially pivotal when dealing with datasets associated with sensitive MHND contexts. *Discourse genre* encompasses the specific communication patterns and linguistic structures used in conversations, therapy sessions, or support interactions related to MHND. These genres vary based on the cognitive and linguistic influences imposed on a speaker [29]. For instance, therapeutic dialogue or narrative retelling are two examples of discourse genres, each requiring distinct linguistic analyses.

Our observation highlights that the interpretation of the term *discourse genre* fluctuates depending on the context of analysis. For instance, in datasets aimed at assessing language disorders, it is frequently denoted by the term itself [29]. In the context of aphasia *discourse genres* [89] or *speech tasks* are commonly used. Bipolar disorder takes the form of *speech tasks* [55], while in cases of depression, it pertains to *speaking patterns* [36] or *tasks* [323]. In the context of dementia, it is linked to *stimuli tasks* [100], *discourse types* [161] or *language production contexts* [137].

In response, we chose to adopt the term *discourse genre* because it serves as a convenient umbrella for referencing the diverse communication patterns used in constructing speech datasets for MHND. However, it is essential to note that in certain cases, such as content derived from online media, pinpointing a specific discourse genre may not be feasible.

In the context of speech MHND datasets, we propose a desideratum that underscores the importance of providing evidence of validity for the intended

use of the data. This involves substantiating that the chosen discourse genres within the dataset are relevant and appropriate for the specified context [185]. For example, in aphasia analysis, narrative tasks may constitute a relevant discourse genre [342, 89], while in depression research, the distinction between discourse genres like reading tasks and interview tasks could be significant [36, 322].

It is essential to note that this desideratum specifically applies to datasets dealing with particularly sensitive MHND information. By establishing evidence of validity, dataset creators ensure that the discourse genres chosen are suitable for the intended context and align with ethical considerations for MHND data. This rigorous validation process helps in maintaining the integrity of the dataset, promoting responsible use in AI applications tailored for MHND support and analysis.

4.3.3 Methodology

In our investigation, we overview existing literature, trying to encompass a broad spectrum of MHND as well as several methods to acquire speech data. Our study includes depression, anxiety,[6] dementia, Alzheimer's, bipolar disorder, somatic symptoms, stress, and Parkinson's. As data sources, we identify personal interaction with participants, online media, and crowdsourcing. For what concerns the genre of the discourse, we include clinical interviews (both unstructured and semi-structured), spontaneous reading, spontaneous speech, speech tasks, and semantic fluency tests.[7]

Selection Criteria

We surveyed relevant literature following a non-systematic approach. We searched titles and abstracts using common keywords related to MHND and speech datasets. First, we explored scientific venues that we consider related to the topic.[8] Then, we extended our search through the Google Scholar search engine. Finally, we examined datasets in the DementiaBank platform,[9] a well-

[6] In literature, anxiety is often mentioned without clear delineation of whether it pertains to a symptom or a syndrome. Since our focus is on MHND, our attention is directed toward datasets that explicitly outline the assessment criteria or scales employed for evaluating anxiety. As a result, we omitted the dataset from [53] due to insufficient information provided regarding the diagnosis of anxiety. Regarding the DAIC-WoZ dataset [97], it was included only for depression analysis since the available data contain only labels for this task.

[7] For detailed definitions of each speech type, please refer to Appendix D.1.

[8] Further details can be found in Appendix D.2.

[9] `https://dementia.talkbank.org/access/`

known repository for datasets concerning MHND, and retrieved publications corresponding to the datasets we are interested in.

Inspired by [170], we then exclude any paper that does not meet all the following criteria: (1) they target patients with neurological disorders; (2) they record speeches by audio or video; (3) they function as public research resources for other researchers; (4) they are peer-reviewed publications. The 36 papers thus selected constitute the final corpus on which we will perform our analysis.

Checklist

To facilitate a structured analysis, we develop a checklist encompassing key considerations reflecting our desiderata. The following checklist is organised into distinct elements and serves as a guide to evaluate the presence or absence of best practices in the reviewed papers.

C1) **Informed Consent.** The authors provide information on sharing informed consent details with participants.

C2) **Data Storage and Security.** The authors report data storage and security measures.

C3) **Data Privacy.** The authors report information on data privacy measures such as data anonymisation, removal of personal information, and naming conventions of data.

C4) **Accountability.** The authors report information about ethical board approval.

C5) **Fairness, Bias, and Diversity.** The authors provide information about data collection diversity and specific considerations for MHND data, including Alzheimer's and Parkinson's, in particular:

- C5-a Speakers distribution per class.
- C5-b Age distribution of the speakers. Authors are required to disclose the average age of each participant group along with at least one additional statistic, such as standard deviation. Alternatively, they should provide the number of participants within specific age ranges, adhering to the inclusion criteria for participant selection.
- C5-c Gender/sex distribution of the speakers.

C5-d Educational level of the speakers.

C5-e Language of the speakers.

C5-f Ethnicity of the speakers.

C5-g **Parkinson's only**. Distribution of Unified Parkinson's Disease Rating Scale (UPDRS) and Hoehn and Yahr Scale (H&Y) among participants.[10]

C5-h Biases awareness. The authors comment on the possible presence of bias in the dataset.

C5-i Diversity collection measures. If the authors explicitly mention if they adopted diversity collection measures.

C6) **Data Quality, Validation, and Maintainability.**

C6-a The authors provide information about technical details on recording settings such as brand and version of the microphone, sampling frequency in Hz or kHz, loudness setting in dB, distance of the microphone from the speaker, and environmental conditions during recording.

C6-b There is information on data maintainability, namely if the platform where data can be accessed allows versioning and issues reporting.

C6-c *Clinical Domain.* There is information on annotation and validation by domain experts, such as the presence of domain experts in the loop.

C6-d *Clinical Domain.* Authors specify the criteria used for diagnosis, regardless of whether it is done by a clinician or through self-assessment questionnaires.[11]

C6-e *Clinical Domain.* **Alzheimer's only**. There are details about the cognitive-linguistic battery. Please refer to Appendix D.1 for the definition of cognitive-linguistic battery.

[10]Please, refer to Appendix D.1 for details about UPDRS and H&Y scales.

[11]When clinicians solely conduct diagnosis, we expect authors to furnish information akin to the following examples: "*In contrast to previous work, all participants met DSM-4 or DSM-5 criteria for major depression as determined by diagnostic interview*" [72]; "*The BDD patients were evaluated and diagnosed by an experienced psychiatrist. All of the patients satisfied both ICD-11 (International Classification of Diseases, 11th Revision) and DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) criteria*" [182]. When the diagnosis is performed through self-assessment questionnaires we expect authors to furnish information akin to the following examples: "*All MDD patients received a structured Mini-International Neuropsychiatric Interview (MINI) that met the diagnostic criteria for major depression of the Diagnostic and Statistical Manual of Mental Disorders (DSM) based on the DSM-IV*" [36]; "*These questions are designed based on DSM-IV and other depression scales such as the Hamilton depression rating scale (HDRS)*" [100].

C7) ***Clinical Domain*** **- Discourse Genres.**

C7-a The authors provide information about the discourse genres participants performed.

C7-b The authors provide motivations for choosing specific discourse types.

4.3.4 Analysis

We present an overview of the 36 surveyed datasets in Table 4.1, detailing their discourse genre, the identified target issues, the sources of speech content, and illustrating their compliance with our checklist. For the sake of brevity, in the rest of this Section, we will abbreviate target issues and types of sources as explained in Appendix A.

Most datasets provide information about speakers and age distribution (C5-a: 28, C5-b: 25), the language of the speakers (C5-e: 24) and discourse genre performed by the participants (C7-a: 28). More than half of the selected works report information about informed consent (C1: 20), accountability (C4: 21), the gender distribution of the speakers (C5-c: 23), presence of domain experts in the loop (C6-c: 17) and reasons for selecting specific discourse genres (C7-b: 17). However, the remaining items are less prominently addressed.

This result indicates a general awareness among the researchers of the importance of reporting the speakers' distribution across different dimensions. However, the speakers' ethnicity is seldom reported. About half of the examined papers do not report on informed consent and accountability. This could be problematic for a number of reasons. First of all, it suggests that these aspects are not sufficiently emphasized in the community. Moreover, it undermines the stability of datasets in case future data policies prevent the use of these data. Likewise, the low coverage in data storage and security (C2: 9) and data privacy (C3: 10) suggests a tendency to overlook measures for the safeguarding of participants' sensitive information. Finally, we observe low coverage in data quality, validation, and maintainability (C6-a: 12, C6-b: 8, C6-d: 8), which could hamper the widespread utilisation of this data.

In the remainder of this section, we analyse our data along three dimensions: (i) the target issues, (ii) the discourse genre and (iii) the sources contributing to the dataset's content. The first and second dimensions aim to chart the community effort concerning specific target issues and discourse genres. The third one intends to highlight potential systematic shortcomings of specific speech sources.

For example, datasets created through crowdsourcing efforts may systematically lack information about recording settings due to the inherent constraints of the collection method. Understanding such patterns is crucial in identifying potential pitfalls of specific content sources.

Dataset	Discourse Genre	Target Issue	Source	C1	C2	C3	C4	C5-a	C5-b	C5-c	C5-d	C5-e	C5-f	C5-g	C5-h	C5-i	C6-a	C6-b	C6-c	C6-d	C6-e	C7-a	C7-b
Zou et al. (2023) [387]	CI	DP	PI	•		•	•	•	•	•		•			•		•	•	•	•	-	•	•
Tlachac et al. (2021) [331]	SR, SS	DP, AX	CS	•	•	•	•									•		•			-	•	•
Dogrucu et al. (2020) [75]	SR	DP	CS	•	•											•					-		
Harati et al. (2020) [106]	CI	DP	PI				•	•	•	•		•							•		-	•	
Lin et al. (2023) [176]	CI	DP, AX	PI	•				•	•	•		•			•	•					-	•	
Tlachac et al. (2022) [330]	CI	DP	CS	•	•	•	•	•	•		•	•	•		•	•		•			-	•	
Tao et al. (2023) [322]	CI, SR	DP	PI	•		•	•	•	•	•	•	•			•			•	•	•	-	•	•
Cai et al. (2022) [36]	CI	DP	PI	•	•	•	•	•	•	•		•					•	•	•	•	-	•	•
Song et al. (2023) [310]	SS	DP, AX	CS					•	•	•		•				•					-		
Tasnim et al. (2022) [323]	ST	DP, AX	CS				•	•	•	•	•	•			•	•			•	•	-	•	•
Gratch et al. (2014) [97]	CI	DP	PI	•		•						•									-	•	•
Shen et al. (2022) [298]	CI	DP	PI	•				•				•			•			•			-	•	
Dibeklioglu et al. (2018) [72]	CI	DP	PI					•	•			•							•	•	-	•	•
Guo et al. (2021) [100]	CI	DP	PI	•			•	•	•	•	•	•					•		•	•	-	•	
Alghowinem et al. (2018) [2]	CI	DP	PI	•			•	•	•			•			•		•		•		-	•	•
Yoon et al. (2022) [374]	SS	DP	OM			•		•		•											-		
Kempler et al. (1987) [137]	SS, ST	DM	PI					•	•		•								•		-	•	•
Kurtz et al. (2023) [156]	ST	DM	PI					•	•	•					•						-	•	
Lanzi et al. (2023) [161]	SR	AL	PI	•	•		•	•	•	•	•	•	•			•	•	•			•	•	•
Luz et al. (2021) [187]	SS	AL	PI		•		•	•	•	•					•			•				•	•
Ivanova et al. (2022) [122]	SR	AL	PI	•			•	•	•	•	•						•		•			•	•
Novikova and Balagopalan (2020) [231]	SS	AL	PI					•															
Becker (1994) [18]	CI	AL	PI	•				•	•	•	•	•							•				
CiftÅŸci et al. (2018) [55]	CI	BD	PI					•	•	•	•	•			•				•	•	-	•	
Wang et al. (2020) [351]	CI	BD	PI	•			•	•	•	•									•	•	-		•
Llamocca et al. (2021) [182]	CI	BD	PI	•		•	•	•				•							•		-		
Gideon et al. (2016) [92]	CI, SS	BD	PI		•		•	•									•		•		-	•	
Qian et al. (2023) [260]	ST	SSD	PI	•	•		•	•	•	•		•			•		•				-	•	
Chaptoukaev et al. (2023) [40]	ST	S	PI	•	•	•	•		•	•	•	•			•		•				-	•	•
Jaiswal et al. (2020) [123]	ST	S	PI							•	•	•					•				-	•	•
Hansen and Bou-Ghazale (1997) [105]	CI, ST	S	PI						•	•											-	•	
Chen et al. (2022) [47]	ST	S	PI			•	•					•									-	•	
Zhang et al. (2022) [381]	SFT	P	PI	•			•	•	•	•	•	•		•							-	•	
Sakar et al. (2013) [284]	SR	P	PI				•	•	•	•							•		•		-	•	•
Orozco-Arroyave et al. (2014) [237]	SR, SS, ST	P	PI	•			•	•	•	•		•		•	•		•		•		-	•	•
Correia et al. (2018) [58]	SS	P, DP	OM									•									-		
Total	36	36	36	20	9	10	21	28	25	23	12	24	2	2	13	7	12	8	17	8	1	28	17

Table 4.1: *Checklist for Mental and Neurological Health Speech Datasets.*

	DP	AL	AX	BD	P	S	DM	SSD
No. Datasets	17	5	4	4	4	4	2	1
C1	11	3	2	2	2	1	0	1
C2	4	2	1	1	0	1	0	1
C3	7	0	1	1	0	2	0	0
C4	9	3	2	3	3	2	0	1
C5-a	13	5	3	4	3	0	2	1
C5-b	11	4	3	2	3	2	2	1
C5-c	9	4	3	2	3	3	1	1
C5-d	4	3	1	1	1	2	1	0
C5-e	14	2	3	2	3	3	0	1
C5-f	1	1	0	0	0	0	0	0
C5-g	-	-	-	-	2	-	-	-
C5-h	7	1	2	1	1	2	1	1
C5-i	6	1	4	0	0	0	0	0
C6-a	4	2	0	1	2	0	0	1
C6-b	6	2	1	0	0	1	0	0
C6-c	8	2	1	4	2	0	1	0
C6-d	5	0	0	2	0	0	0	0
C6-e	-	1	-	-	-	-	-	-
C7-a	13	3	3	2	3	4	2	1
C7-b	8	3	2	1	2	2	1	0

Table 4.2: *Number of datasets that address each checklist item grouped by target issue.*

Dimension 1: Target Issue

In Table 4.2, we group the datasets according to their target issues to provide an overview of their compliance with the items in our checklist. Datasets encompassing multiple target issues are considered multiple times, once per issue. In our analysis, we will follow the same order represented in the table.

Depression — Among the 17 datasets concerning depression (**DP**), we observe that 11 incorporate details on informed consent (C1), 4 implement robust data storage and security measures (C2), 7 disclose information on data privacy measures (C3), and 9 report on accountability (C4). Regarding fairness, bias, and diversity aspects, we found that 13 datasets provide information on the distribution of speakers by class (C5-a), 11 inform regarding the age distribution of speakers (C5-b), 9 present the gender/sex distribution of speakers (C5-c), and 4 outline the education level of speakers (C5-d). Moreover, 14 datasets include

information on different languages of speakers (C5-e), but only one elaborates on distinct speaker ethnicity for a specific language (C5-f). Lastly, 7 datasets exhibit awareness of potential bias in the dataset (C5-h), and 6 incorporate diversity collection measures (C5-i). Regarding aspects of data quality, validation, and maintainability, only 4 datasets provide information about the recording setting (C6-a), and 6 adhere to the principles of maintainability (C6-b). 8 datasets involve the presence of domain experts in the loop (C6-c), while the others automatically validate observations based on scores obtained from participants in depression self-assessment questionnaires. Among those involving domain experts, 5 out of 8 specify the type of clinical diagnosis manual adopted (C6-d). Finally, 13 datasets provide information regarding the speech tasks required of participants (C7-a), and only 8 of them justify that choice (C7-b).

We argue that the limited awareness of potential biases in the datasets and the relatively low adoption of diversity collection measures (C5) should raise concerns about the inclusivity and representativeness of the research. Moreover, the frequent absence of experts in the loop highlights a potential oversight, risking a lack of nuanced understanding and interpretation. The limited information on recording settings and adherence to transparency and maintainability principles further suggests a need for the scientific community to identify best practices in ensuring data quality (C6). The uneven distribution in providing justifications for the choices made about discourse genre suggests a need for more thorough and transparent methodological reporting. In conclusion, we believe that further reflection is needed on the integration of ethical considerations in depression datasets.

Alzheimer's — According to Table 4.2 (column **AL**), the majority of datasets provide information on the distribution of speakers by class (C5-a), details about speaker age (C5-b) and speaker gender/sex (C5-c). Furthermore, around half of the datasets cover informed consent (C1), accountability (C4), information about speaker educational levels (C5-d), and discourse genre (C7-a and C7-b). However, only two datasets feature aspects such as data security (C2), language of the speakers (C5-e) and data quality, validation, and maintainability (C6-a, C6-b, C6-c). Similarly, a single dataset provides documentation regarding different ethnicities of speakers for a specific language (C5-f), bias awareness (C5-h), diversity collection measures (C5-i), and cognitive-linguistic batteries (C6-e). Notably, there is no mention of the adoption of data privacy measures (C3).

Overall, the analysis of Alzheimer's datasets reveals consistent attention to certain ethical dimensions, such as speaker distribution and demographic details. However, critical aspects like data security and privacy, as well as measures addressing bias and diversity, demand increased consideration in future research.

Anxiety — From Table 4.2 (column **AX**), we observe that aspects such as informed consent (C1), speaker distribution by class (C5-a), speaker age (C5-b), speaker gender/sex (C5-c), different languages of speakers (C5-e), and information on discourse genres (C7-a, C7-b) are well-represented, with more than half of datasets reporting information in these areas. Moreover, all datasets report information about the adoption of diversity collection measures (C5-i).

However, other critical ethical dimensions, such as data storage and security (C2), data privacy (C3), accountability (C4), educational level of participants (C5-d), and different ethnicities of speakers for a specific language (C5-f), are less frequently addressed, being present in at most two out of the seven collected datasets. Lastly, all aspects concerning data validity and maintainability (C6) are scarcely reported or entirely omitted.

In conclusion, while certain ethical considerations receive ample attention in anxiety disorder datasets, there remains a significant need for increased emphasis on crucial aspects such as data security, privacy, data quality, maintainability, and the involvement of domain experts to ensure a more robust ethical framework across the analyzed papers.

Bipolar Disoder — We collect 4 datasets addressing Bipolar Disorder (**BD**). We observe that the majority of them provide details on accountability (C4), speaker distribution (C5-a) and involve domain experts for data validation/annotation (C6-c). However, only half of these datasets include information on the clinical diagnosis reference manual (C6-d). Furthermore, aspects such as data storage and security (C2), data privacy (C3), speaker education level (C5-d), and bias awareness (C5-h) are scarcely addressed. Moreover, when it comes to data quality, only one dataset includes information about the recording setting (C6-a). Lastly, crucial aspects like the inclusion of speakers of different languages or the acknowledgment of different ethnicities in the dataset (C5-e, C5-f), the adoption of diversity collection measures (C5-i), and maintainability are not documented.

In conclusion, while some datasets provide valuable insights into speaker distribution and involve domain experts, there is a conspicuous lack of documentation on essential ethical considerations and aspects related to data quality.

Parkinson's — We collect 4 datasets related to Parkinson's (P). Three of them provide information on accountability (C4), as well as the distribution of speakers by class, age, gender/sex, and education level (C5-a, C5-b, C5-c) and discourse genres performed by speakers (C7-a). However, only two datasets offer justifications for their choice of discourse genres. Additionally, only half of the analyzed papers report details on informed consent (C1), the language of the speakers (C5-e), UPDRS and H&Y rating scales (C5-g), recording settings (C6-a), and validation by domain experts (C6-c). In contrast, only one dataset contains information on the education level of speakers (C5-d) or bias awareness (C5-h). Notably, none of the papers discusses data storage and security (C2), data privacy (C3), ethnicity of speakers (C5-d and C5-f), diversity collection measures (C5-i), or maintainability (C6-b).

Despite the limited availability of Parkinson's datasets for analysis, our results suggest that there are notable gaps in addressing crucial ethical considerations. Future research in Parkinson's studies should prioritize aspects like data security, privacy, diversity measures, and maintainability to ensure a more comprehensive and ethically robust framework for data collection in the context of Parkinson's disease. Moreover, we argue that more attention has to be paid to report recording setting details and UPDRS and H&Y rating scales.

Stress — Table 4.2 (column **S**) reports statistics about datasets targeting Stress. All analysed datasets provide information on the discourse genre performed by speakers (C7-a). About half of them include details on the age and gender/sex distribution of speakers (C5-b, C5-c), measures of data privacy (C3), accountability (C4), level of speaker education (C5-d), biases awareness (C5-h), and rationale for choosing discourse genres (C7-b). Nonetheless, only one dataset includes information on crucial aspects such as informed consent (C1) and data storage and security (C2). Furthermore, no dataset addresses speaker distribution by class (C5-a), speaker's ethnicity (C5-f), and diversity collection measures (C5-i). Lastly, the majority of collected datasets do not report details about data quality, validation and maintainability (C6).

Overall, there is a general lack of information on key ethical considerations. In particular, aspects like data storage security, speaker distribution by class, and diversity collection measures are insufficiently addressed.

Somatic Symptoms Disorder and Dementia — A relatively low number of datasets address Somatic Symptoms Disorder (SSM) and Dementia (DM). We ob-

serve a notable lack of information regarding informed consent (C1), data storage and security (C2), data privacy (C3), and accountability (C4). In the context of collection diversity and documentation tailored to address mental health diseases, we observe a good coverage of speakers' distribution (C5-a) and their age (C5-b). Nonetheless, other relevant speakers' characteristics, like gender/sex distribution (C5-c), education level (C5-d), language (C5-e), ethnicity (C5-f), and diversity collection measures (C5-i), are scarcely addressed. Likewise, almost no dataset reports information concerning data quality, validation, and maintainability (C6)

Overall, despite the relatively small sample of datasets, our observations highlight a lack of documentation regarding the data collection process, the diversity of involved speakers, and, most importantly, its inherent quality and validation.

Dimension 2: Discourse Genre

Table 4.3 reports the number of datasets that employ specific discourse genres and provides a comprehensive overview of compliance with the items in our checklist. It is worth noting that some datasets encompass multiple discourse genres. Thus, the same dataset may appear multiple times when counting occurrences.

Clinical Interviews — A clinical interview serves as a fundamental discourse genre within the domain of healthcare, facilitating an in-depth interaction between patients and clinicians. In essence, it is a structured or semi-structured conversation designed to gather comprehensive information about a patient's medical history, current symptoms, and psychosocial factors. Unlike casual conversations, clinical interviews are purposeful and goal-oriented, aiming to elicit specific information crucial for diagnosis, treatment planning, and therapeutic interventions.

In our study, Clinical Interviews (**CI**) are the most prevalent discourse genre. A good portion of the datasets report informed consent (C1), highlighting a strong commitment to participant autonomy and transparency in research procedures. Additionally, there is significant attention given to fairness, bias, and diversity (C5), indicating a concerted effort to mitigate potential biases and ensure equitable representation within the datasets. However, only two datasets provide considerations about diversity collection measures (C5-i).

Accountability (C4) emerges as another important aspect, demonstrating adherence to institutional review board protocols and ethical oversight. Moreover, it is worth noting that clinical domain-related items, such as the presence of domain experts in the loop (C6-c), and reporting discourse genre information (C7-a), are

	CI	ST	SS	SR	SFT
No. Datasets	17	9	9	7	1
C1	12	3	2	6	1
C2	3	2	3	3	0
C3	6	2	2	2	0
C4	10	5	4	6	1
C5-a	15	5	7	5	1
C5-b	13	7	4	5	1
C5-c	10	7	4	5	1
C5-d	5	4	1	3	1
C5-e	14	6	3	3	1
C5-f	1	0	0	1	0
C5-g	0	1	1	1	1
C5-h	7	5	2	2	0
C5-i	2	1	2	3	0
C6-a	5	4	2	4	0
C6-b	5	0	2	3	0
C6-c	12	3	3	4	0
C6-d	7	1	0	1	0
C6-e	0	0	0	1	0
C7-a	14	9	5	6	1
C7-b	7	5	4	6	0

Table 4.3: *Number of datasets that address each checklist item grouped by discourse genre.*

present in more than half of the datasets, even if information on clinical diagnosis is present only in seven datasets (C6-d).

Given the direct relationship between clinicians and patients during data collection through clinical interviews, it is reassuring to observe significant attention toward informed consent and accountability. However, we argue that greater attention should be paid to data storage and security as well as data quality, validation, and maintainability to mitigate risks and ensure the protection of sensitive participant information. Lastly, given the intimate nature of these interactions, where participants often divulge personal and sensitive information to clinicians, it becomes imperative to prioritize data protection and anonymization, especially when utilizing and sharing it for research endeavors.

Speech Tasks — Speech Tasks (**ST**) can encompass a broad spectrum of activities, ranging from simple repetition of phonemes to complex narrative retellings.

The nature of these tasks is tailored to the objectives of the study, aiming to elicit specific linguistic features or behaviors indicative of underlying cognitive or physiological conditions.

As we can observe in Table 4.3, these datasets exhibit similar ethical considerations to CI datasets, albeit with some distinctions. While fairness, bias, and diversity considerations feature prominently (C5-b, C5-c, and C5-e), suggesting efforts to promote inclusivity and mitigate biases in dataset composition, there is a lack of documentation concerning informed consent (C1), accountability (C4), and biases awareness (C5-h). However, clinical domain-related items are less emphasised in the context of data validation (C6).

It is encouraging that all the analysed datasets describe the speech tasks performed by participants (C7-a), even if around half of them also provide reasons for choosing them (C7-b). Unfortunately, there is scarce attention towards data storage and security (C2), diversity collection measures (C5-i), and data quality and maintainability (C6-a, C6-b). The aspect of privacy (C3) is often not addressed properly too, but it is important to consider that in this context participants would probably be less inclined to divulge sensitive information.

Spontaneous Speech — Spontaneous speech (**SS**) addresses the unscripted, fluid expression of language that occurs in everyday interactions. Unlike rehearsed or scripted speech, spontaneous speech tasks prompt individuals to articulate their thoughts, emotions, and experiences spontaneously, mirroring the dynamic nature of real-world conversations. Whether engaging in casual conversations, narrating personal anecdotes, or responding to open-ended prompts, participants generate spontaneous speech that reflects their linguistic abilities, cognitive processes, and sociolinguistic awareness.

Table 4.3 highlights that nearly half of the datasets provide details about accountability (C4), the distribution of participants by age and gender/sex (C5-b, C5-c), speaker interaction and corresponding motivation of such a choice (C7-a, C7-b). Additionally, a low portion of the datasets addresses data storage and security (C2), privacy (C3), educational level (C5-d), language (C5-e), ethnicity (C5-f), biases awareness (C5-h), and diversity collection measures (C5-i). Furthermore, there is a lack of focus on data quality, validation, and maintainability (C6). Lastly, it is particularly concerning the neglect of data privacy (anonymization), despite its significance in a scenario where people speaking spontaneously may report personal and sensitive information.

Spontaneous Reading — Spontaneous reading (**SR**) targets the natural flow and comprehension of written text among individuals. Unlike scripted readings or recitations, spontaneous reading tasks entail participants engaging with text material without prior preparation or rehearsal.

From Table 4.3 we observe that nearly all datasets address informed consent (C1) and accountability (C4). Furthermore, the majority of the datasets address fairness, bias, and diversity (C5-a, C5-b, C5-c) and data quality, validation, and maintainability (C6-a, C6-b, C6-c) Additionally, most datasets detail the types of spontaneous reading tasks assigned to participants and provide the rationale for their selection (C7-a, C7-b), a crucial aspect for understanding study objectives and participant characteristics relevant to syndrome detection or classification. However, few works discuss data storage and security (C2) or data privacy (C3), although participants may be less inclined to divulge sensitive information in this context. While participants may not disclose sensitive information, vigilance in these areas is crucial for maintaining data integrity and reliability. Similarly, information regarding the educational level (C5-d) and language (C5-e) of participants is lacking, despite its relevance to interpreting results from reading tasks fairly. Moreover, few datasets provide details about participant ethnicity (C5-f), bias awareness (C5-h), and diversity collection measures (C5-i).

Semantic Fluency Tests — Semantic fluency tests (**SFT**) evaluate individuals' cognitive functioning like their semantic memory and language abilities. The primary objective is to measure the individual's capacity to access and retrieve information from their semantic memory stores spontaneously. In these tests, participants are typically asked to generate words belonging to a specific semantic category within a limited timeframe, such as naming animals or fruits [381].

In our survey, SFT is limited to the dataset provided by [381]. Drawing definite about the comprehensive coverage of aspects outlined in our checklist on a single dataset is challenging. However, we can infer that this dataset prioritises certain aspects. Specifically, attention is given to informed consent (C1), accountability (C4), fairness, bias, and diversity considerations (C5). Additionally, the dataset includes a description of the discourse genre (C7-a). Nonetheless, no information about data evaluation, validity, and maintainability (C6) is reported.

	CS	OM
No. Datasets	5	2
C1	3	0
C2	3	0
C3	2	1
C4	3	0
C5-a	3	1
C5-b	3	0
C5-c	2	1
C5-d	2	0
C5-e	3	1
C5-f	1	0
C5-g	0	0
C5-h	2	0
C5-i	5	0
C6-a	0	0
C6-b	2	0
C6-c	1	0
C6-d	1	0
C6-e	0	0
C7-a	3	0
C7-b	2	0

Table 4.4: *Datasets addressing each checklist item grouped by data collection method.*

Dimension 3: Source of Content

We observe that data is typically collected either via personal interactions with participants (PI), indirect methods such as crowdsourcing (CS), or by annotating media already available online (OM). We posit that CS or OM data collection prevents compliance with specific aspects of the checklist, whereas PI does not. Accordingly, in the analysis that follows, we focus on CS and OM. Table 4.3 offers a detailed view.

Crowdsourcing — Crowdsourcing (**CS**) involves outsourcing content-production tasks from a distributed pool of participants, who are tasked to engage in various activities, such as completing surveys, performing cognitive tasks, collecting social media and/or sensor data, or providing subjective feedback on their health status.

In our survey, we consider 5 datasets gathered through crowdsourcing. We

find it noteworthy that all of these datasets incorporate diversity collection measures (C5-i). This implementation is particularly seamless in crowdsourced data collection, as participants typically provide personal information themselves.

Information concerning recording settings (C6-a) is absent in all the papers. This is understandable, considering that in crowdsourcing, participants often utilise their personal laptops or mobile devices, resulting in very diverse settings. In most cases, details about the involvement of domain experts and clinical diagnoses (C6-c, C6-d) are lacking. This suggests that data annotation is largely automated, possibly relying on self-assessment scores.

Online Media Annotation — Two of the datasets we analysed concerned cases where the source videos and audios were already publicly available on YouTube. The authors of the datasets scraped such material, annotated it, and redistributed it. In these cases, information regarding informed consent (C1), data storage and security (C2), and accountability (C3) are not provided. The point about data storage and security is understandable since the user content is not produced or managed by the authors of the dataset. However, we believe the issue of informed consent and accountability deserves attention. We are afraid that this is an inherent limitation of online media annotation, because collecting informed consent from the original content creators or the approval of ethical boards could be challenging, if not outright unfeasible. Likewise, for the same reason, diversity collection measures (C5-i) are not reported in either case.

4.3.5 Findings

Our survey identifies key areas for improvement in mental health and neurological disorders resource creation and use. About half of the examined papers do not report on informed consent and accountability. Other crucial aspects, such as data storage, privacy, quality, validation, and maintainability, have even lower coverage. These limitations are not always distributed equally among the domains of interest. For instance, the absence of validation from experts in the field predominantly impacts datasets related to Anxiety and Stress, as well as those gathered through discourse genres outside of clinical interviews. Another example concerns datasets collected through online media annotations, which typically lack information on informed consent, data storage and security, as well as accountability measures.

On a positive note, our survey has also demonstrated a rich and varied set

of valuable resources for MHND research. Such resources are pivotal in driving progress toward reliable and trustworthy AI systems with a positive impact on society, and this survey may serve as a reference to the novice embarking on such an endeavor. However, we believe it is crucial that such resources are used with increased awareness of potential pitfalls. To this end, we compile a list of desiderata and a checklist for collecting and documenting datasets for mental health and neurological disorders. The checklist is intended not only as a tool to analyse existing literature but also, perhaps more importantly, as a guideline for a more principled approach to future resource design, implementation, maintainance and use in this field.

We are aware that the application of such guidelines may pose some challenges. One of them would be addressing the trade-offs between separate items, such as data privacy and data access [15]. In this regard, researchers agree that each study needs to consider the benefits of both items at tension with each other, and find an equilibrium - e.g., balance the need for seamless exchange of information for improved healthcare outcomes with the strong obligation to protect the patient's privacy [52]; or balance the risks of a user's privacy loss with the benefits of delivering them a tailored experiences and targeted marketing from shared data [192]. Even more so, in light of these challenges, we wish for our work to stimulate a closer collaboration between computer science and clinical domains for a sounder and more mutually beneficial resource development.

4.4 TRI-DEP: A Trimodal Comparative Framework for Depression Detection Using Speech, Text, and EEG

This section presents our multimodal depression detection research, addressing critical methodological gaps that have constrained progress in this domain [233]. This work was conducted in collaboration with Annisaa Fitri Nurfidausi and Paolo Torroni. As established in the chapter introduction (Section 4.1), current automated depression detection systems suffer from three interconnected limitations: most studies combine only two modalities, overlooking comprehensive trimodal integration; the field lacks systematic comparisons between handcrafted features and modern pretrained embeddings, particularly brain-pretrained models for EEG; and inconsistent evaluation protocols have made it difficult to assess true generalization performance to unseen patients.

Key Contribution: Our work systematically addresses these limitations through a complete exploration of feature representations and modelling strategies across EEG, speech, and text modalities. We compare handcrafted features against embeddings from large, pre-trained models, including, for the first time in multimodal depression detection research, brain-pretrained architectures for EEG alongside multilingual speech encoders and transformer-based language models. We establish rigorous experimental protocols using stratified subject-level cross-validation with fixed splits across all experiments, ensuring genuine assessment of model generalisation.[a] Our systematic exploration demonstrates that carefully designed trimodal models outperform both unimodal and bimodal configurations.

[a] Code available at `https://github.com/annisaafitrinn/tri-dep`

Beyond identifying the best-performing architecture, we introduce an experimental framework that serves as a reference setup for future research, enabling systematic evaluation of new fusion strategies or additional modalities while maintaining methodological rigour. To ensure reproducibility and support further advances in the field, we release all data splits, preprocessing scripts, and the complete implementation.

4.4.1 Methodology

This section describes our experimental approach for trimodal depression detection. We detail the dataset characteristics, preprocessing pipelines for each modality, feature extraction methods encompassing both handcrafted descriptors and pretrained embeddings, modality-specific neural architectures, and multimodal fusion strategies employed throughout our study.

Data

This study employs the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) [36], which provides: (1) 5-minute resting-state EEG recorded with a 128-channel HydroCel Geodesic Sensor Net at 250 Hz, and (2) audio from structured clinical interviews. For each subject, the interview audio consists of $R = 29$ separate recordings (question−answer items) whose durations vary across and within subjects (the total interview time is approximately 25 minutes per subject). Since MODMA does not include text transcriptions from clinical interviews, we generate automatic transcriptions using speech-to-text models. The dataset comprises individuals diagnosed with MDD, recruited from Lanzhou University Second Hospital, and Healthy Control (HC) obtained via public advertising; MDD diagnoses were confirmed by licensed psychiatrists. In this study, we retain only subjects who participated in both EEG and interview recordings, resulting in a filtered cohort of 38 subjects. Table 4.5 summarises demographic information across groups and protocols. Additional details are available in [36].

Table 4.5: *Participant demographics in the MODMA dataset. M: Male, F: Female, HC: Healthy Control, and MDD: Major Depressive Disorder.*

Modality	Total	MDD (M/F)	HC (M/F)	Age MDD/HC)
128-ch EEG	53	24 (13/11)	29 (20/9)	16–56 / 18–55
Speech	52	23 (16/7)	29 (20/9)	16–56 / 18–55

MODMA's compliance with ethical guidelines is detailed in Section 4.3 (Table 4.1, row 8), demonstrating satisfaction of core criteria (informed consent, ethical approval, clinical validation). While certain desiderata (e.g., educational level, ethnicity) are not fully met, essential requirements are satisfied.

Many studies lack clarity in data splitting [259, 375, 109], where segment-level splits can leak information by placing recordings from the same subject in both

training and test sets, yielding inflated performance. To avoid this, we use stratified 5-fold subject-level cross-validation with consistent splits across experiments. We also release these splits on our companion website[12] to ensure reproducibility and fair comparison.

To address the lack of transcriptions in the MODMA dataset, we employ WhisperX [14] to generate text for each subject's 29 recordings, without further post-processing. We acknowledge ASR systems are known to exhibit reduced accuracy on clinical speech [210, 132], potentially introducing transcription errors; however, ground-truth transcriptions are not available for verification.

Experimental Pipeline Design

We design a unified pipeline for multimodal depression detection with EEG, speech, and text. For EEG, we adopt two processing branches: a 29-channel, 250 Hz, 10 s segmentation setup, consistent with prior work [259, 375, 109, 139], and a 19-channel, 200 Hz, 5 s segmentation setup replicating the preprocessing used in CBraMod for the MUMTAZ depression dataset [223] (which contains resting-state EEG data obtained from 33 depressed patients and 30 healthy controls). For CBraMod, we evaluated both the original pre-trained version and the model fine-tuned on MUMTAZ, as described in the official documentation[13] , and found the latter consistently superior. Therefore, throughout this work, we refer to CBraMod as the MUMTAZ-fine-tuned model. Speech recordings are resampled to 16 kHz, denoised, and segmented into 5 s windows with 50% overlap, while text is used directly from raw Chinese transcriptions.

Feature extraction combines handcrafted descriptors (EEG statistics, spectral power, entropy; speech MFCCs with/without prosody) with embeddings from large pre-trained models. For EEG, we employ both the Large Brain Model (LaBraM), trained on ∼2,500 hours of EEG from 20 datasets, and CBraMod, a patch-based masked reconstruction model. For speech, we use XLSR-53, a multilingual wav2vec 2.0 encoder, and Chinese HuBERT Large, trained on 10k hours of WenetSpeech. For text, we use Chinese BERT Base, MacBERT, XLNet, and MPNet Multilingual. Segment-level representations are encoded with a combination of CNNs, LSTMs, and/or GRUs (with/without attention) and fused using decision-level strategies.

[12] https://annisaafitrinn.github.io/tri-dep/

[13] https://github.com/wjq-learning/CBraMod/blob/main/

Data Preprocessing

The preprocessing stage serves multiple objectives, including cleaning and structuring the raw data, as well as preparing it for multimodal analysis. One key objective is the segmentation of the input into smaller units that can be more effectively processed by the models. We denote with $\mathbf{S}_{\text{EEG}}$, $\mathbf{S}_{\text{SPEECH}}$, and $\mathbf{S}_{\text{TEXT}}$ the number of segments obtained after preprocessing for each input modality.

EEG — For handcrafted features and LaBraM, we follow prior work [259, 375, 109, 139], which comprises retaining $C = 29$ channels[14], applying a $0.5{-}50\,\text{Hz}$ bandpass filter with a 50 Hz notch, and average re-referencing. Recordings are segmented into 10 s windows; at 250 Hz, each window contains $T = 250 \times 10 = 2500$ samples. Thus, a recording of length L seconds produces $S_{\text{EEG}} = L/10$ windows (e.g., $S_{\text{EEG}} = 30$ for a 5-min recording), represented as $\mathbf{X}^{(1)}_{\text{EEG}} \in \mathbb{R}^{S_{\text{EEG}} \times C \times T}$.

For CBraMod, we use the version pretrained on the MUMTAZ depression dataset, thereby replicating its preprocessing. Signals are resampled to 200 Hz, bandpass filtered ($0.3{-}75\,\text{Hz}$) with a 50 Hz notch, and reduced to $C = 19$ channels[14]. Recordings are segmented into 5 s windows; at 200 Hz, each window contains $T = 200 \times 5 = 1000$ samples. A recording of length L seconds thus yields $S_{\text{EEG}} = L/5$ windows (e.g., $S_{\text{EEG}} = 60$ for a 5-min recording). Each window is further divided into $P = 5$ non-overlapping patches of $T_{\text{patch}} = 200$ samples, resulting in $\mathbf{X}^{(2)}_{\text{EEG}} \in \mathbb{R}^{S_{\text{EEG}} \times C \times P \times T_{\text{patch}}}$.

Speech — Audio recordings are resampled from 44 kHz to 16 kHz, converted to mono PCM, amplitude-normalized to $[-1, 1]$, silence-trimmed, and denoised with a median filter [91]. Each signal is segmented into overlapping windows of length $w = 5\,\text{s}$ with hop size $h = 2.5\,\text{s}$ (50% overlap). At a sampling rate of 16 kHz, each segment contains $T_{\text{seg}} = 80{,}000$ samples and each hop $T_{\text{hop}} = 40{,}000$ samples.

For a recording of duration L seconds (post-trimming), the number of segments is $S_{\text{SPEECH}} = \left\lfloor (L - w)/h \right\rfloor + 1$ for $L \geq w$, while recordings shorter than w are retained as a single segment. The segmented waveform is represented as $\mathbf{X}_{\text{SPEECH}} \in \mathbb{R}^{S_{\text{SPEECH}} \times T_{\text{seg}}}$, where each row corresponds to one waveform segment. Each subject has $R = 29$ interview recordings; after windowing, recording r yields $S^{(r)}_{\text{SPEECH}}$ segments $\mathbf{X}^{(r)}_{\text{SPEECH}} \in \mathbb{R}^{S^{(r)}_{\text{SPEECH}} \times T_{\text{seg}}}$. The subject-level speech representation is the concatenation along the segment axis: $\mathbf{X}_{\text{SPEECH}} = \left[\mathbf{X}^{(1)}_{\text{SPEECH}}; \ldots; \mathbf{X}^{(R)}_{\text{SPEECH}}\right]$, with a total of $S_{\text{SPEECH}} = \sum_{r=1}^{R} S^{(r)}_{\text{SPEECH}}$ segments.

[14] Full list available on our companion website.

Text — Each recording has a single transcript. After tokenisation, the subject-level text representation is the concatenation of all transcript representations, $\mathbf{X}_{\text{TEXT}} = \left[\mathbf{X}^{(1)}_{\text{TEXT}}; \ldots; \mathbf{X}^{(R)}_{\text{TEXT}}\right]$.

Feature Extraction

EEG — *Handcrafted features.* For each segment $\mathbf{X}^{(1)}_{\text{EEG}} \in \mathbb{R}^{C\times T}$ we extract $F = 10$ handcrafted descriptors per channel (statistical, spectral, entropy), yielding $\mathbf{X}_{\text{HAND}} \in \mathbb{R}^{S\times C\times F}$.

Pre-trained models. We further extract embeddings from LaBraM and CBraMod. LaBraM operates on $\mathbf{X}^{(1)}_{\text{EEG}}$ and maps each segment to a $D = 200$-dimensional embedding, producing $\mathbf{X}_{\text{LaBraM}} \in \mathbb{R}^{S\times D}$. CBraMod operates on $\mathbf{X}^{(2)}_{\text{EEG}}$, where each 5 s segment is patch-encoded and then averaged across channels and patches to form $D = 200$-dimensional embeddings, resulting in $\mathbf{X}_{\text{CBraMod}} \in \mathbb{R}^{S\times D}$.

Remark. After feature extraction, the raw temporal dimension (T or T_{patch}) is no longer present, as each segment is reduced to a fixed-size representation of dimension F (handcrafted) or D (embeddings). For subject-level modelling, features from all recordings of the same subject are stacked to form the final subject representation.

Speech — From each waveform segment, we compute two handcrafted variants, namely MFCCs (40 coefficients) and Prosody + MFCCs (46 features: 40 MFCCs plus energy, F_0, RMS energy, pause rate, phonation time, speech rate). We also extract segment embeddings with XLSR-53 and Chinese HuBERT Large. Segment-level features are stacked per recording and then concatenated across the 29 recordings of each subject to form the subject-level representation. The exact tensor shapes (per recording and subject-level) are summarised in Table 4.6. After feature extraction, the raw sample length is no longer present; each segment is represented by a fixed-size vector (40/46/768/1024 dimensions).

Table 4.6: *Shapes of speech feature matrices per recording r and at the subject level.*

Feature name	Per-recording shape	Subject-level shape
$\mathbf{X}_{\text{MFCC}}$	$\mathbb{R}^{S^{(r)}_{\text{SPEECH}}\times 40}$	$\mathbb{R}^{S_{\text{SPEECH}}\times 40}$
$\mathbf{X}_{\text{PROSODY+MFCC}}$	$\mathbb{R}^{S^{(r)}_{\text{SPEECH}}\times 46}$	$\mathbb{R}^{S_{\text{SPEECH}}\times 46}$
$\mathbf{X}_{\text{XLSR}}$	$\mathbb{R}^{S^{(r)}_{\text{SPEECH}}\times 1024}$	$\mathbb{R}^{S_{\text{SPEECH}}\times 1024}$
$\mathbf{X}_{\text{HuBERT}}$	$\mathbb{R}^{S^{(r)}_{\text{SPEECH}}\times 768}$	$\mathbb{R}^{S_{\text{SPEECH}}\times 768}$

Text — Each recording has a single transcript, which we encode with a pretrained language model (BERT, MacBERT, XLNet, or MPNet) to obtain one $D = 768$-dimensional embedding per recording; for a subject with $R = 29$ recordings, stacking these yields $\mathbf{X}_{\text{BERT}}, \mathbf{X}_{\text{MacBERT}}, \mathbf{X}_{\text{XLNet}}, \mathbf{X}_{\text{MPNet}} \in \mathbb{R}^{R \times D}$ (with $R = 29$, $D = 768$).

Baselines

We re-implement two multimodal baselines for depression detection that use standard image architectures on EEG and speech *2D-spectrograms (Spec2D)*: DenseNet-121 [375] and Vision Transformer (ViT) [258]. These studies are among the few that explore EEG−speech multimodality in this task and report promising results. In our experiments, we retain their model architectures but apply our own subject-level cross-validation splits for consistency, making results not directly comparable to the original works. Additional implementation details are provided on our companion website.

Architectures

We assess several modality-tailored architectures. We also experiment with multimodality, combining predictions from the best-performing feature-model pair in each modality in a late fusion fashion.

To keep notation light, we use $\mathbf{F}$ to denote the *generic feature matrix* per modality, either handcrafted features or embeddings from pretrained models. Concretely, $\mathbf{F}_{\text{EEG}}$ (EEG), $\mathbf{F}^{(r)}_{\text{SPEECH}}$ (speech, per recording r), and $\mathbf{F}_{\text{TEXT}}$ (text, subject-level).

EEG — We consider three detection modules: *CNN+LSTM*, *Gated Recurrent Unit (GRU)+Attention*, and *CNN*. The CNN+LSTM branch uses two 1D convolutions (kernel size 3, padding 1) with dropout to capture local temporal patterns, followed by a 2-layer LSTM for sequence modelling. *GRU+Attention* uses a 2-layer GRU with an attention mechanism that weights hidden states to form a subject-level summary. *CNN* retains only the convolutional layers followed by max-pooling and an MLP head, dispensing with recurrence entirely. All three modules consume $\mathbf{F}_{\text{EEG}}$ and produce a latent representation $\mathbf{H}_{\text{EEG}}$; an MLP head outputs y_{EEG}.

Speech — A shallow CNN extracts segment-level features from each recording. These are reduced to a single fixed-size vector $\mathbf{F}^{(r)}_{\text{SPEECH}} \in \mathbb{R}^d$ using one of three

encoders: (i) max pooling, (ii) GRU with attention, or (iii) BiGRU with attention (the latter extending the GRU+Attn design with bidirectional recurrence). The resulting $R = 29$ vectors are stacked into the subject-level matrix $\mathbf{F}_{\text{SPEECH}} \in \mathbb{R}^{R \times d}$. This sequence is then processed by a subject-level classifier—either an LSTM or a CNN with max-pooling—to produce the subject-level representation $\mathbf{H}_{\text{SPEECH}}$, which is fed to an MLP head to obtain the final prediction y_{SPEECH}.

Text — $\mathbf{F}_{\text{TEXT}}$ denotes the subject-level text features (Sec. 4.4.1). A detection module (LSTM or CNN) transforms $\mathbf{F}_{\text{TEXT}}$ into $\mathbf{H}_{\text{TEXT}}$, and an MLP head outputs y_{TEXT}.

Multimodal Fusion — We select, for each modality, the best-performing feature-modelpair and fuse their predictions via decision-level (late) fusion. This design choice is motivated by modularity and interpretability: each modality is trained independently with its own optimal architecture, and fusion operates on output predictions, making it straightforward to add or remove modalities without retraining. This property is particularly valuable in clinical settings, where not all modalities may be available for every patient. We consider three schemes: *weighted averaging* – compute weighted combination of modality probabilities where weights sum to one, then predict the class with highest weighted probability; *Bayesian fusion* – convert modality-specific posteriors to likelihood ratios, combine them with predefined weights, and map back to a posterior; *majority voting* – each modality independently predicts a class label, and the final prediction is the class receiving the most votes, with ties resolved arbitrarily at 0.5.

4.4.2 Experimental Setup

We adopt stratified 5-fold cross-validation with fixed subject splits to ensure balanced and comparable experiments and prevent data leakage. Models are trained with cross-entropy loss and softmax output, with hyperparameters tuned manually. All implementation details are provided on our companion website. Given the limited sample size (38 subjects, 3–4 validation subjects per fold), early stopping criteria based on validation metrics are unreliable, since a single misclassification shifts validation accuracy by 25–33%. To avoid overfitting to noisy validation signals, all models are trained for a fixed number of epochs, determined per configuration through preliminary experiments on the validation set. At the end of training, the final model state is used for inference. This ensures a consistent and reproducible training protocol across all modalities and configurations.

4.4.3 Results

In this section, we report the performance of all experimental categories: baseline re-implementations, unimodal models, and our proposed multimodal architectures. F1-scores are reported as mean ± standard deviation across folds. Table 4.7 presents the baselines and unimodal models, including the best-performing model for each set of features per modality. Table 4.8 reports the performance of baseline models and multimodal fusion strategies, highlighting the best configuration within each category and the overall best-performing model. Further results are available on our companion website.

Unimodal — Table 4.7 reports the performance of baseline and unimodal models. Among EEG features, CBraMod embeddings with a CNN classifier achieve the best result (F1 = 0.446), confirming the benefit of fine-tuning on a depression-related corpus, though all EEG configurations remained below 0.5 in terms of macro F1-score. For speech, Chinese HuBERT embeddings coupled with a BiGRU encoder and CNN+MaxPool detection module achieve the strongest performance (F1 = 0.809), while handcrafted MFCC and prosodic features yield lower scores, indicating that deep speech embeddings capture richer information. In the text modality, Chinese MacBERT with an LSTM reach the top result (F1 = 0.784). Overall, speech provides the most informative single modality, closely followed by text, while EEG remained less predictive in isolation.

Category	Features	Detection Module	F1
Baselines (Speech+EEG)	$\mathbf{X}_{\text{Spec2D}}$	ViT	0.560 ± 0.190
	$\mathbf{X}_{\text{Spec2D}}$	DenseNet-121	0.586 ± 0.240
EEG	$\mathbf{X}_{\text{HAND}}$	CNN+LSTM	0.418 ± 0.096
	$\mathbf{X}_{\text{LaBraM}}$	GRU+Attn	0.417 ± 0.089
	$\mathbf{X}_{\text{CBraMod}}$	**CNN**	**0.446 ± 0.106**
Speech	$\mathbf{X}_{\text{MFCC}}$	CNN+MaxPool → LSTM	0.553 ± 0.143
	$\mathbf{X}_{\text{Prosody+MFCC}}$	CNN+BiGRU+Attn → LSTM	0.655 ± 0.182
	$\mathbf{X}_{\text{XLSR}}$	CNN+MaxPool → CNN+MaxPool	0.774 ± 0.143
	$\mathbf{X}_{\text{HuBERT}}$	**CNN+BiGRU+Attn → CNN+MaxPool**	**0.809 ± 0.081**
Text	$\mathbf{X}_{\text{MPNet}}$	CNN	0.725 ± 0.244
	$\mathbf{X}_{\text{BERT}}$	CNN	0.720 ± 0.122
	$\mathbf{X}_{\text{XLNet}}$	LSTM	0.665 ± 0.088
	$\mathbf{X}_{\text{MacBERT}}$	**LSTM**	**0.784 ± 0.114**

Table 4.7: *Results of baselines and unimodal models (macro-F1, mean ± std across 5 folds). The Features column indicates the input representation (handcrafted or learned); the Detection Module column indicates the classifier trained on top. For speech, the detection module comprises a segment-level encoder and a subject-level classifier, separated by →. In **bold**, the best performing configuration per modality.*

Multimodal — Table 4.8 compares the baselines with different fusion strategies. Simple baselines such as ViT and DenseNet-121 reach F1-scores around 0.56. Fusion weights are optimised via grid search (step = 0.05) over all modality combinations. The trimodal configurations achieve the highest macro F1-score for both weighted averaging (0.864) and Bayesian fusion (0.864), outperforming the best unimodal model (speech, F1 = 0.809) by approximately 5.5 percentage points. Among bimodal configurations, Speech+Text consistently perform best across all fusion methods, with weighted averaging achieving F1 = 0.839 and notably low variance (std = 0.059). These results confirm the complementarity of modalities: while speech and text individually provide strong unimodal performance, integrating all three modalities yields the strongest overall result.

Category	Configuration	F1-score
Baselines	ViT	0.560 ± 0.190
	DenseNet-121	**0.586 ± 0.240**
Weighted Averaging	EEG + Speech (0.05 : 0.95)	0.809 ± 0.081
	EEG + Text (0.15 : 0.85)	0.786 ± 0.116
	Speech + Text (0.45 : 0.55)	0.839 ± 0.059
	EEG + Speech + Text (0.05 : 0.35 : 0.60)	**0.864 ± 0.095**
Bayesian Fusion	EEG + Speech (0.05 : 0.95)	0.809 ± 0.081
	EEG + Text (0.05 : 0.95)	0.784 ± 0.114
	Speech + Text (0.20 : 0.80)	0.863 ± 0.140
	EEG + Speech + Text (0.05 : 0.35 : 0.65)	**0.864 ± 0.095**
Majority Voting	EEG + Speech	0.557 ± 0.135
	EEG + Text	0.500 ± 0.078
	Speech + Text	**0.809 ± 0.081**
	EEG + Speech + Text	0.800 ± 0.200

Table 4.8: *Baseline and multimodal fusion results (F1-score, mean ± std across 5 folds). In **bold**, the best performing configuration per category. The overall best are additionally underlined. Fusion weights were optimised via grid search (step = 0.05).*

Experimental Framework for Multimodal Depression Detection — Building on this systematic exploration of feature extraction methods, neural architectures, and fusion strategies, we propose an experimental framework for multimodal depression detection, illustrated in Figure 4.1. The framework selects the best-performing predictors for each modality: $\mathbf{X}_{\text{CBraMod}}$ processed with a CNN for EEG, $\mathbf{X}_{\text{HuBERT}}$ processed with a CNN+BiGRU+Attn encoder and CNN+MaxPool classifier for speech, and $\mathbf{X}_{\text{MacBERT}}$ processed with an LSTM for text. These modality-specific pipelines are then combined through alternative

decision-level fusion strategies. Our best-performing configurations use either weighted averaging or Bayesian fusion across the three modalities with optimized weights, both achieving a macro-F1 of 0.864 averaged across 5 folds.

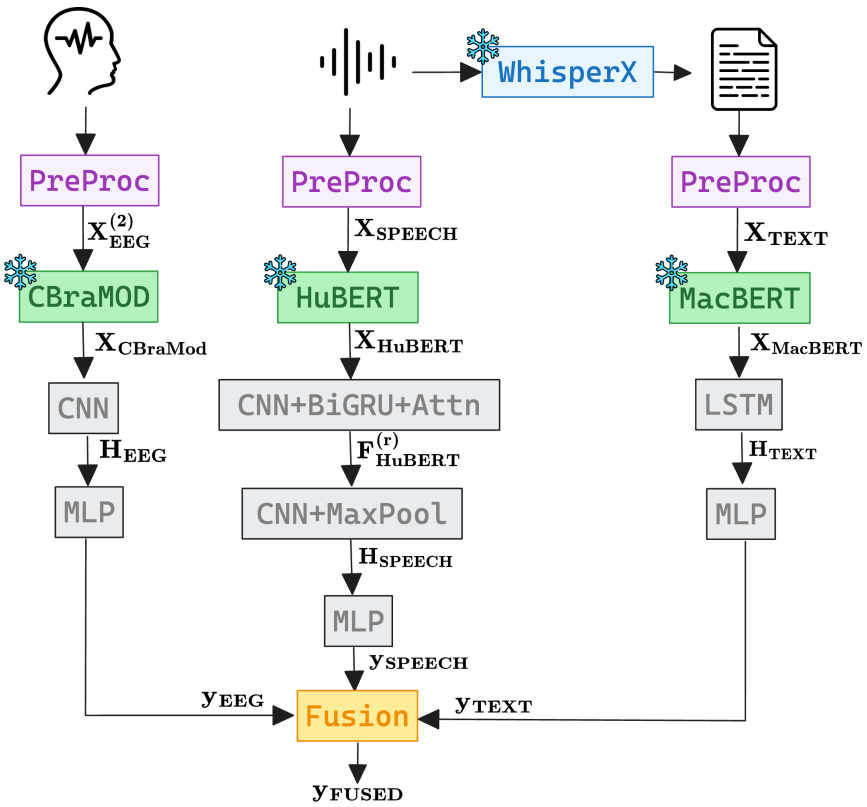


Figure 4.1: *Experimental Framework for Multimodal Depression Detection.*

To assess whether performance differences are statistically significant, we conduct pairwise tests across all 253 pairs of the 23 configurations reported in Tables 4.7–4.8 (11 best unimodal feature–model pairs + 12 fusion configurations).

We employ two complementary non-parametric tests, following [74]: McNemar's exact test, which evaluates whether two classifiers make significantly different errors on the same subjects (N=38), and the Wilcoxon signed-rank test, which assesses whether per-fold macro-F1 differences are consistent across folds (n=5) without assuming normality.

Table 4.9 reports results for 12 selected comparisons that address three key questions: whether each trimodal fusion method outperforms the best unimodal and best bimodal configurations, whether the best bimodal (weighted averaging fusion between speech and text) outperforms individual modalities, and how the three modalities rank against each other. "Best" refers to the best-performing unimodal configuration per modality (bold rows in Table 4.7): Best

Text = $\mathbf{X}_{\text{MacBERT}}$+LSTM, Best Speech = $\mathbf{X}_{\text{HuBERT}}$+BiGRU+CNN+MaxPool, Best EEG = $\mathbf{X}_{\text{CBraMod}}$+CNN.

Model A	**Model B**	**ΔF1**	**McNemar p**	**Wilcoxon p**
WA: EEG+Speech+Text	Best Text	+0.081	0.6476	0.2379
BF: EEG+Speech+Text	Best Text	+0.081	0.7257	0.2379
MV: EEG+Speech+Text	Best Text	+0.016	1.0000	1.0000
WA: EEG+Speech+Text	Best Speech	+0.055	0.8765	0.6286
BF: EEG+Speech+Text	Best Speech	+0.055	0.7673	0.6286
WA: Speech+Text	Best Text	+0.055	0.9297	0.3906
WA: Speech+Text	Best Speech	+0.030	1.0000	1.0000
WA: EEG+Speech+Text	WA: Speech+Text	+0.025	1.0000	1.0000
BF: EEG+Speech+Text	WA: Speech+Text	+0.025	1.0000	1.0000
Best Text	Best Speech	-0.025	1.0000	0.6286
Best Text	Best EEG	+0.337	0.1472	0.2379
Best Speech	Best EEG	+0.363	0.1002	0.1832

Table 4.9: *Pairwise statistical significance tests for 12 selected comparisons from the 253 total pairs across the 23 configurations reported in Tables 4.7–4.8. ΔF1 is the difference in mean macro-F1 between Model A and Model B (positive values indicate Model A outperforms Model B). No comparison reaches significance at* α=0.05. *WA = Weighted Averaging, BF = Bayesian Fusion.*

With only N=38 subjects and n=5 folds, no comparison reaches statistical significance at α=0.05; notably, the minimum achievable p-value for the Wilcoxon test with 5 paired observations is 0.0625. Furthermore, although fusion does not yield statistically significant improvements over the best unimodal model at this sample size, it consistently increases macro-F1 and improves stability: weighted averaging Speech+Text achieves a standard deviation of 0.059, lower than both text alone (0.114) and the trimodal configuration (0.095), indicating more stable predictions across different patient subgroups.

The framework serves as a reference setup for future experiments, enabling systematic evaluation of new fusion strategies or additional modalities. The limited statistical power observed in our significance analysis is consistent with the broader landscape of clinical affective computing, where datasets typically comprise 30–50 subjects and pairwise significance is rarely achievable. Similar power limitations have been reported in related multimodal depression detection studies, and remain an open challenge for the field. Despite this constraint, the consistent direction of improvement across multiple fusion methods and modality combinations provides converging evidence for the benefit of multimodal integration. To the best of our knowledge, our trimodal configuration with weighted averaging

fusion represents state-of-the-art performance in multimodal depression detection.

4.4.4 Findings

We address key limitations in multimodal depression detection by adopting subject-level stratified cross-validation and exploring EEG-based representations in combination with speech and text. Our experiments compare handcrafted features with deep representations from large pretrained models, consistently showing the superiority of the latter across all three modalities. In the unimodal setting, speech emerges as the most predictive modality, closely followed by text, while EEG remains the least discriminative in isolation.

In the multimodal setting, decision-level fusion consistently improves over unimodal models, with trimodal configurations achieving the strongest results. A notable finding is the asymmetric role of EEG: despite its weak unimodal performance, it provides complementary information that benefits trimodal fusion, but only in the presence of text. Bimodal speech+text fusion offers a particularly favourable trade-off between performance and stability, with weighted averaging achieving F1 = 0.839 and the lowest cross-fold variance of any configuration (std = 0.059).

Pairwise statistical significance tests across all reported configurations reveal that no comparison reaches significance at α=0.05, which is expected given the small sample size (N=38) and the inherent power limitations of non-parametric tests with 5 folds. This is consistent with the broader landscape of clinical affective computing, where small cohorts remain the norm. Despite this constraint, the consistent direction of improvement across multiple fusion methods and modality combinations provides converging evidence for the benefit of multimodal integration.

Beyond identifying the best-performing architecture, we introduce an experimental framework that serves as a reference setup for future research, enabling systematic evaluation of new fusion strategies or additional modalities while maintaining methodological rigour. While decision-level fusion ensures modularity, it limits the ability to capture cross-modal interactions; our framework provides a natural starting point for investigating early and intermediate fusion strategies. The MODMA dataset's reliance on automated transcription via WhisperX, rather than ground-truth transcriptions, represents a limitation, as ASR systems exhibit elevated word error rates on clinical speech [210, 132]. However, our text modality

performance suggests this impact remained moderate. Future work could investigate transcription quality systematically or explore ASR systems specifically trained on clinical populations.

To ensure reproducibility and support further advances in the field, we release all data splits, preprocessing scripts, and the complete implementation.

4.5 Investigating the Effectiveness of Explainability Methods in Parkinson's Detection from Speech

This section presents our systematic evaluation of explainability methods for speech-based clinical detection, using Parkinson's disease as a case study. This work was conducted in collaboration with Francesco Paissan, Paolo Torroni, Mirco Ravanelli, and Cem Subakan [195].[15] As emphasised in the chapter introduction (Section 4.1), model interpretability represents a critical yet underexplored challenge in clinical AI, particularly for speech and audio modalities. While text-based models can highlight words or phrases that domain experts readily interpret, speech explanations face a fundamentally different challenge: identifying influential spectrogram regions may capture what the model attends to, but whether these technical representations translate into clinically interpretable insights remains unclear. This uncertainty raises fundamental questions: can current explainability methods provide insights that healthcare professionals can interpret and act upon, or do inherent limitations in explaining acoustic features constrain their practical value for clinical deployment?

Key Contribution: We present a comprehensive evaluation of perturbation-based and gradient-based explanation methods, including Saliency, SmoothGrad, Integrated Gradients, Guided GradCAM, Guided Backpropagation, and Guided SHAP.[a] We evaluate these methods across multiple dimensions: quantitative faithfulness metrics, overlap analysis, auxiliary classification experiments, and selective metrics that penalise explanations replicating the entire input. Our findings reveal a troubling disconnect: although explanations achieve high technical faithfulness — correctly identifying regions that influence classifier decisions — they often fail to provide insights that domain experts can readily interpret or act upon.

[a] Code available at: `https://github.com/helemanc/parkinsons-speech-xai`

This result has profound implications for the broader field of speech-based and multimodal AI systems that include speech, suggesting that before we can build truly explainable multimodal systems, we must first solve the more fundamental problem of making speech explanations interpretable to domain experts, partic-

[15] © 2025 IEEE. Reprinted, with permission, from [195].

ularly healthcare professionals in clinical settings. To support further research in this direction, we release our complete implementation and provide extensive visualisations through our companion website.[16]

4.5.1 Methodology

Our methodology to compare and evaluate XAI methods consists of (i) obtaining attributions and saliency maps using mainstream interpretability techniques, (ii) evaluating the explanatory power of such saliency maps (and their combination) quantitatively via a range of metrics defined in the literature, and (iii) evaluate the information conveyed by saliency maps for the Parkinson's Disease (PD) detection task from an auxiliary classifier.

Transforming Waveforms for Gradient-Based Attributions

Many Self-Supervised Learning (SSL)-based pre-trained models for audio operate directly on the raw waveform. This is the case also for our PD detection model, HuBERT, which is described in detail in Section 4.5.2. In this case, explaining the model predictions in the classifier's input domain (i.e. on the waveform X_w) results in explanations that are hard to interpret visually. We transform the waveforms into a time-frequency representation X_f using the short-time Fourier transform (STFT). This transformation allows us to compute *attributions* (A) on the spectrogram of the input audio via saliency-based interpretability techniques. Before inputting the audio to the model, we convert the spectrogram back to the time domain using the inverse short-time Fourier transform (ISTFT) and the phase information of the original sample X_w. We apply the following interpretability techniques to generate these maps: Saliency [302], SmoothGrad [305], Integrated Gradients (IG) [319], Guided GradCAM [293], Guided Backpropagation [312], and Guided SHAP [186].

Quantitative Analysis through Explainability Metrics

To quantitatively evaluate the quality of the saliency maps, we employ several explainability metrics. Specifically, we adopt metrics previously used in the L-MAC [242], LMAC-ZS [241] and LMAC-TD [194] studies. We use Average Increase (AI), which measures the percentage of samples for which we observe an increase in the classifier's confidence for the interpretation with respect to the

[16] `https://helemanc.github.io/parkinsons-speech-xai/`

input sample, Average Decrease (AD), which measures the confidence drop when masking the input with the mask, and Average Gain (AG), similar to AI. Beyond these metrics, we use the Faithfulness (FF) metric defined in the L2I paper [245] and the input fidelity (Fid-In) metric defined in the PIQ paper [243]. We also use the Sparseness (SPS) [38] and Complexity (COMP) [26] metrics to evaluate the conciseness of the explanations. We invite the reader to refer to the cited papers for further details.

Exploring Overlap in Explainability Methods

A reasonable expectation when explaining a classifier is that good explanations are aligned with the classifier's predictions, regardless of the process that generated them. To verify this hypothesis, we investigate the overlap among different attribution methods to assess whether combining them provides further insights into the model's decisive process. Given a dataset element (X_f, y), we extract two saliency maps A_1, A_2 using two different explanation techniques (e.g. Saliency and Smoothgrad) and combine them via intersection and union. An increase in explainability metrics through overlapping attributions would suggest that combining methods captures more comprehensive feature representations.

Classification Using Saliency Maps and Selective Metrics

Finally, to examine whether the saliency maps derived from each attribution method highlight information pertinent to distinguishing between people with PD and HCs, we train a classifier on the generated explanations. This test is similar to RemOve And Retrain [115]; in our case, however, we measure the amount of information in the explanations in a single iteration. Additionally, we introduce a new set of metrics, called *Selective Metrics*, that scales the standard classification metrics depending on the explanation selectiveness. These metrics penalise explanations that replicate the input audio without selectively highlighting the portions of the input relevant to the classification.

To define selective metrics, we combine the classification performance (e.g. accuracy) and the average of the attribution mask. Numerically, for a dataset $\mathcal{D} = \{(X_f, y)_i\}_{i=1}^{N}$ of N elements and metric $M(X_f, y)$ (i.e. accuracy, where X_f is the STFT domain input audio), this computes as:

$$S_{\mathrm{M}}(\mathcal{D}) = \frac{1}{N} \sum_{(X_f, y) \in \mathcal{D}} M(X_f, y)\,(1 - \mathrm{Av}(A)) \tag{4.1}$$

where $\mathrm{Av}(\cdot)$ denotes the mean that is calculated on the normalized attributions A (which is in $[0, 1]$). This adjustment rewards classifiers that achieve high accuracy while focusing on smaller, relevant input parts. This is particularly relevant to facilitate a comparison between attribution strategies.

4.5.2 Experiments

Dataset

Standard PC-GITA (s-PC-GITA) is a dataset consisting of recordings from 100 individuals, split evenly between two groups: 50 people with PD and 50 HC. Each group includes 25 men and 25 women, with the PD group diagnosed by a neurologist, while the HC group shows no signs of PD or other neurodegenerative conditions. Participants range in age from 31 to 86 years, with recordings captured in a soundproof booth at Clínica Noel in Medellín, Colombia. The original recordings were sampled at 44.1 kHz with 16-bit resolution and downsampled to 16 kHz for this study, as in [227] and [157]. We use the same splits as those employed by [157] for inference on the test sets of the 10 folds and present the results averaged across these folds; our replicated results, which align with the original paper, are shown in Table 4.10. As in [157], the speech tasks from both datasets considered in this work are diadochokinetic (DDK) exercises, read sentences, and monologues. Additional dataset details are available in [238].

Model	Accuracy	F1-score
HuBERT Base [117]	81.32 ± 8.06	81.03 ± 8.33
WavLM Base [46]	82.10 ± 7.94	81.90 ± 8.09

Table 4.10: *PD detection results averaged over 10-folds on s-PC-GITA. Mean value and standard deviation are reported.*

PD Detector

In this study, we build upon recent advancements in the field by relying on a study that explores exploiting foundation models and speech enhancement for Parkinson's disease detection from speech in real-world operative conditions [157]. In [157], they use SSL models HuBERT[117] and WavLM[46] as foundational backbones for PD detection, extracting high-level representations from raw waveform inputs through a convolutional and transformer-based encoder. By leveraging

Metric	Accuracy	Selective Accuracy	F1-score	Selective F1-score	Mask Mean	Mask Std
Saliency	0.87 ± 0.12	0.86 ± 0.10	0.87 ± 0.14	0.85 ± 0.13	0.017	0.037
Smoothgrad	0.86 ± 0.11	0.85 ± 0.11	0.86 ± 0.13	0.85 ± 0.12	0.013	0.020
Guided GradCAM	0.78 ± 0.09	0.78 ± 0.09	0.78 ± 0.10	0.78 ± 0.10	0.002	0.010
Guided Backprop	0.87 ± 0.10	0.86 ± 0.10	0.85 ± 0.14	0.84 ± 0.13	0.016	0.035
Integrated Gradients	**0.89 ± 0.08**	**0.88 ± 0.08**	**0.89 ± 0.09**	**0.88 ± 0.09**	0.008	0.023
Gradient SHAP	0.84 ± 0.13	0.83 ± 0.13	0.85 ± 0.13	0.84 ± 0.13	0.007	0.019
Original Spectrograms	0.82 ± 0.12	-	0.85 ± 0.10	-	-	-

Table 4.11: *Results of PD detection using saliency maps.*

pre-trained layers optimized through dynamic weighted summation and an attention pooling head, the model captures discriminative features essential for PD detection, refining these representations with fully connected layers to enhance task-specific accuracy. Following this work, we replicate their results using only the s-PC-GITA datasets and apply explainability techniques to the model.

Saliency Maps Classifier

We employ a CNN14 classifier [149] to evaluate saliency maps, taking log-mel spectrograms as input. This classifier is trained for a binary classification task with the following hyperparameters: batch size of 32, learning rate of 0.002, and 50 training epochs.

To train and evaluate the classifier, we construct a dataset from the masks computed on the original test set. For each fold, this dataset is split into training, validation, and test sets with a ratio of 70/15/15, ensuring that the label distribution is stratified to maintain balanced classes across the splits. For comparison, we also train the CNN14 classifier on the original spectrograms from the test set to evaluate performance differences between saliency map-based inputs and the unmodified spectrograms. The CNN14 classifier training is done using the SpeechBrain 1.0 toolkit [271].

4.5.3 Results

As shown in Table 4.10, both the HuBERT Base and WavLM Base models obtain comparable results on PD detection. For this reason, we conduct our experiments on HuBERT in this section. We note that the same experimental protocol and methodology can be applied to interpreting WavLM. Our code is publicly available and can be accessed through our companion website, together with additional spectrogram visualisations.

Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Saliency	74.66 ± 7.06	1.80 ± 2.22	64.09 ± 13.01	0.004 ± 0.003	82.82 ± 14.27	0.69 ± 0.02	11.98 ± 0.09
Smoothgrad	75.10 ± 8.43	1.85 ± 1.86	55.54 ± 13.42	0.004 ± 0.002	81.89 ± 11.76	0.50 ± 0.02	12.51 ± 0.05
Guided GradCAM	64.99 ± 11.71	5.24 ± 6.74	55.85 ± 16.06	0.001 ± 0.001	**82.85 ± 13.83**	**0.83 ± 0.02**	**10.69 ± 0.12**
Guided Backprop	75.21 ± 6.71	1.62 ± 2.06	64.52 ± 12.90	0.005 ± 0.004	82.67 ± 14.03	0.70 ± 0.02	11.94 ± 0.11
Integrated Gradients	**78.83 ± 6.56**	**1.22 ± 1.72**	**69.35 ± 9.20**	**0.013 ± 0.008**	81.93 ± 14.32	0.77 ± 0.01	11.69 ± .00
Gradient SHAP	76.55 ± 7.59	2.26 ± 3.65	67.46 ± 10.60	0.004 ± 0.003	81.93 ± 13.75	0.69 ± 0.01	11.99 ± 0.06

Table 4.12: *PD detection quantitative evaluation results of HuBERT Base model averaged over 10-folds on s-PC-GITA.*

Faithfulness Metrics

In Table 4.12, we present the quantitative evaluation results of the HuBERT-base model averaged over 10 folds on s-PC-GITA. We observe that all the XAI approaches show comparable performance overall. The variability in the metrics can be attributed to the high sensitivity of classifier representations to individual subjects, as evidenced by the high standard deviations in Table 4.10. This subject-specific variability leads to fluctuations in faithfulness metrics across folds comparable to those observed on the classification performance. For AI, AD, AG, and FF, Integrated Gradients performs best, while for Fid-In, SPS, and COMP, Guided Grad-CAM stands out. This also aligns with the results presented in Table 4.11, in which Guided Grad-CAM shows the lowest mask mean, aligning with its superior performance in SPS and COMP. We note that the small FF values are due to the classifier's uncertainty, since we observed small values in the predicted logit outputs. Overall, we see that the explanations are aligned with the classifier, suggesting that masking the audio spectrogram can be an effective way of highlighting the regions of the input audio associated with the predicted label.

Overlap between Explainability Methods

Fig. 4.2 shows a scatter plot illustrating how faithfulness metrics vary with increasing intersection-over-union (IoU) between explanations from different methods, specifically using intersection as the combination strategy. For the results of the overlap based on the union strategy, we do not observe a linear trend with increasing IoU values. However, we observe that for some metrics (e.g. AD), union is a more effective strategy on average. We focused on two metrics, AD and FF, which showed the most variation across methods. A similar trend is observed on other metrics; we invite the reader to refer to the companion website for the results. Overall, the trend suggests that combining attribution strategies is most

effective when the attributions are already well-aligned, supporting our hypothesis that greater mask overlap improves faithfulness metrics.

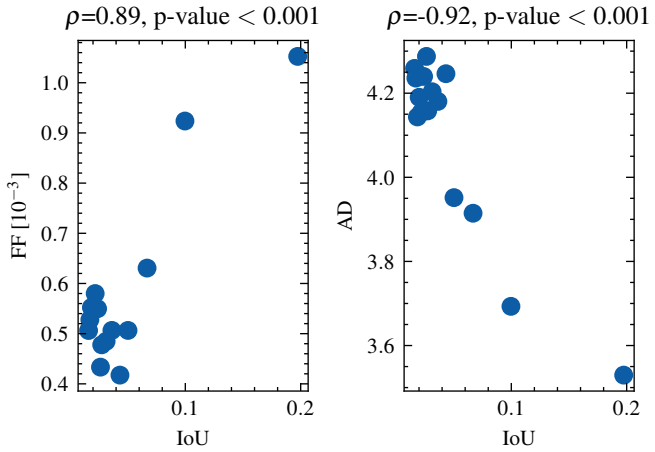


Figure 4.2: *Analysis of the correlation between explanations overlap among interpretability techniques and faithfulness metrics. Combining interpretability techniques is most effective when there is already significant overlap between attribution masks. Figure taken from [195]. © 2025 IEEE.*

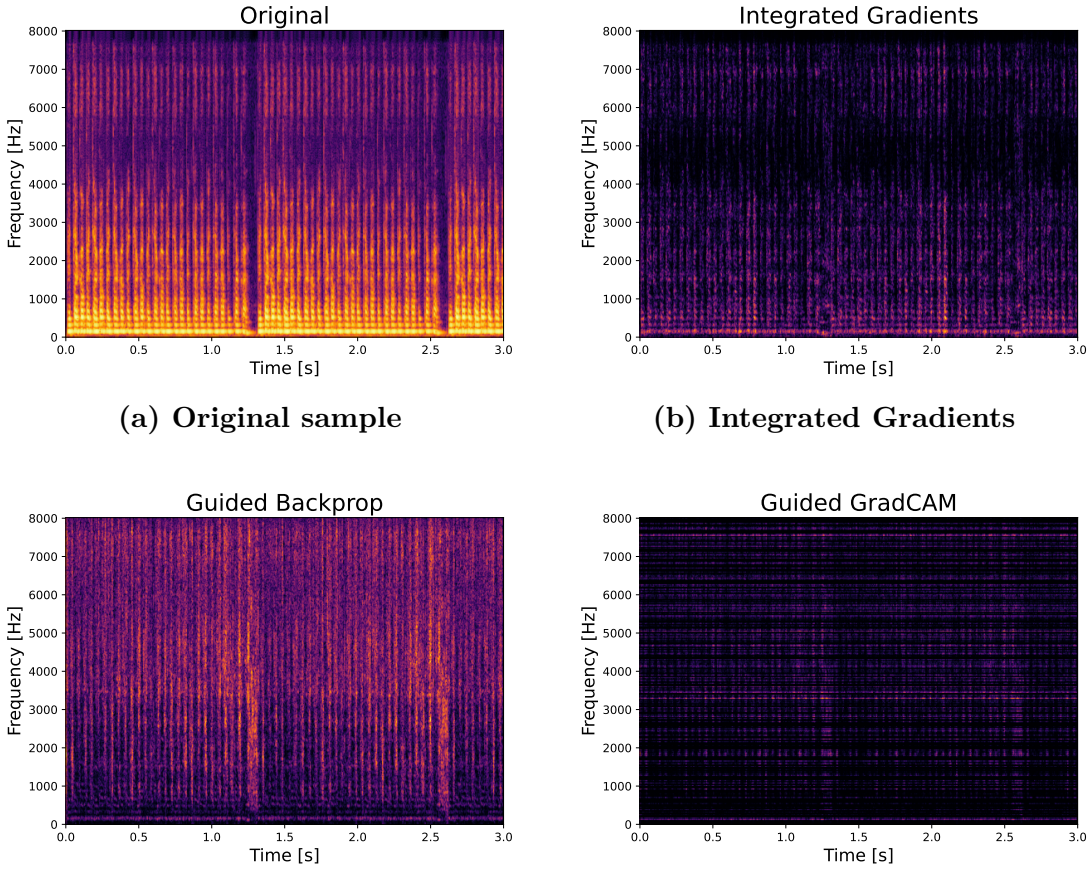
Saliency Maps Classifier

In Table 4.11, we report the classification accuracy and F1-score, together with the corresponding selective metrics. In most cases, training the classifier on the explanations results in higher classification performance with respect to training on the original data. This suggests that training on the explanations provides better generalization capabilities, possibly because by incorporating explanations, the classifier gains reduced context dependency, thus enhancing its generalization capabilities. Among the XAI methods, Integrated Gradients stands out, excelling in faithfulness metrics and achieving the highest selective accuracy, confirming its overall effectiveness compared to other strategies.

Qualitative Analysis

Quantitative results show that explanations align with the classifier outputs and provide helpful information for distinguishing between PD and HC. In Figure 4.3, we present a sample explanation generated with Guided GradCAM, Integrated Gradients and their combination. To assess interpretability, we analysed 10 samples per speech task across four scenarios: correctly and incorrectly classified PD and HC samples. Visualisations are available on our companion website.

In general, we observed that the saliency maps focus on high-frequency regions, potentially reflecting attention to specific phonemes. Nonetheless, these maps are not easily interpretable by humans. We conclude that despite current XAI techniques provide faithful explanations as spectrograms, further research is needed to render explanations more insightful for domain experts. Relevant works in this direction include [101], which relies on additional data modalities.



(a) Original sample **(b) Integrated Gradients**

(c) Guided Backprop **(d) Guided GradCAM**

Figure 4.3: *Explanations generated for a PD sample correctly classified by HuBERT. The explanations highlight different portions of the spectrogram, suggesting that understandability is difficult to achieve in this setting. From left to right, top to bottom: original sample, Integrated Gradients, Guided Backprop, and Guided GradCAM. Figure taken from [195]. © 2025 IEEE.*

4.5.4 Findings

We show that popular post-hoc explanation methods successfully generate faithful explanations for PD detection, establishing a solid technical foundation. The key challenge we identify is translating these faithful explanations into formats that

align with domain experts' perceptual capabilities.

This opens promising avenues for future work. We propose three concrete directions: (1) semantically enriching spectrograms through phoneme discretization to create direct links between saliency maps and established speech biomarkers; (2) developing listenable explanation methods (extending approaches like L-MAC) that present insights in auditory forms matching clinicians' diagnostic processes; and (3) exploring hybrid visualisations combining spectral representations with linguistically meaningful annotations. By building on our validated foundation of faithful explanations, these approaches can advance toward clinically actionable interpretability.

4.6 Conclusion

This chapter presents a comprehensive investigation of multimodal AI for clinical decision support, addressing three interconnected challenges: ethical data collection and curation, effective multimodal fusion, and clinical interpretability. Through three complementary studies, it establishes both the foundational requirements and the current limitations of deploying trustworthy multimodal systems in MHND research.

Section 4.3 examines ethical practices in speech dataset development, using the DAIC-WoZ depression corpus as a case study. It defines a set of desiderata spanning informed consent, privacy, accountability, fairness, data quality, and discourse genre. A systematic review of 36 datasets reveals significant shortcomings: nearly half fail to document informed consent or accountability, while data security and maintainability are often neglected. These findings highlight that ethical integrity is not ancillary but foundational to credible and effective multimodal research.

Section 4.4 demonstrates the advantages of comprehensive multimodal integration for depression detection. Pretrained embeddings consistently outperform handcrafted features across modalities, with brain-pretrained EEG models and large speech encoders providing strong representations. The proposed trimodal architecture achieves 86.4% F1-score, outperforming unimodal and bimodal systems under rigorous subject-level validation that prevents data leakage and ensures genuine generalisation.

While these results confirm the feasibility of high-performing multimodal clinical systems, Section 4.5 exposes a critical barrier to real-world deployment: the gap between model explanations and user interpretability. Gradient-based and perturbation-based explainability methods achieve high technical faithfulness but offer limited epistemic value to clinicians. This limitation extends beyond healthcare to any domain where speech-based or multimodal AI requires transparency.

Together, these findings reveal three key insights. First, ethical data practices are indispensable to trustworthy multimodal systems. Second, performance gains from multimodal integration are meaningful only when obtained under principled validation protocols. Third, explainability must evolve from technically faithful to clinically interpretable methods to enable genuine understanding and trust.

This work provides both conceptual and practical guidance. The dataset cre-

ation checklist and the experimental framework offer concrete tools for ethical, reproducible multimodal research. The released infrastructure, including standardised data splits and preprocessing pipelines, enables fair comparison of future approaches.

Future research should prioritise the development of explainability techniques that connect model behaviour to interpretable, clinically meaningful concepts — potentially through phoneme-level discretisation, listenable explanations, or multimodal scaffolding that leverages text interpretability as an anchor. Addressing this challenge benefits not only healthcare but all high-stakes domains where acoustic information informs decision-making. Furthermore, advancing annotation methodologies that explicitly link paralinguistic cues to diagnostic indicators represents a crucial next step.

Overall, multimodal AI for clinical decision support embodies both opportunity and responsibility. Ethical compromises undermine trust; uninterpretable accuracy risks misuse; biased models threaten equity. Progress demands systems that are simultaneously accurate, interpretable, and ethically grounded. The methodological foundation established in this work demonstrates that such a balance is achievable and essential for the responsible use of speech-based AI in clinical and broader societal contexts.

Link to Research Questions. This chapter primarily addresses **RQ6**, **RQ7**, and **RQ8**, while also contributing to the cross-domain **RQ1** and **RQ3**. Together, these contributions define the ethical, technical, and interpretability foundations for multimodal clinical decision support.

- **RQ1: Multimodal Necessity** — The results confirm that multimodal integration substantially enhances clinical assessment, as speech, text, and EEG capture complementary aspects of mental and neurological conditions that unimodal systems overlook.
- **RQ3: Systematic Solutions** — The proposed experimental frameworks, standardised data splits, and reproducible implementations provide systematic tools for evaluating speech-based multimodal architectures in high-stakes clinical contexts.
- **RQ6: Ethical Framework Development** — Addressed through a comprehensive checklist for dataset creation and ethical auditing, emphasising informed consent, privacy, accountability, and fairness as prerequisites for responsible multimodal dataset curation.
- **RQ7: Trimodal Clinical Assessment** — Empirical evidence demonstrates that combining speech, text, and EEG signals yields significant performance gains over unimodal and bimodal systems, confirming the benefits of holistic multimodal fusion for depression detection.
- **RQ8: Clinical Speech Biomarkers** — The explainability analysis reveals that existing attribution techniques fail to provide clinically meaningful insights, underscoring the need for new interpretability frameworks that connect acoustic evidence to diagnostic constructs and clinician understanding.

CHAPTER 5

Multimodal Musical Version Identification

5.1 Introduction

Music presents a fundamentally different multimodal challenge than the political discourse and clinical assessment examined in previous chapters. While speech and text modalities represent parallel channels of the same content, music involves a more intricate relationship between lyrics and audio. Music is one of the most inherently multimodal forms of human artistic expression, in which meaning emerges from the intricate interplay between audio content and lyrical text. Lyrics constitute a distinct artistic dimension, while audio encompasses separate acoustic elements—such as melody, harmony, and instrumentation—that can vary independently.

This independence clearly appears in everyday music perception and processing. Consider how listeners recognise different versions of the same song, whether a faithful musical version, a dramatic reinterpretation, or a translation into another language. Humans naturally integrate information from both modalities, identifying versions that maintain lyrical content while transforming musical ar-

rangement (a jazz reinterpretation of a pop song), versions that preserve musical elements while translating text (a Spanish version of an English original), and even versions that alter both dimensions while retaining sufficient similarity (a metal band covering a folk ballad in another language). This cognitive integration reveals that effective music understanding requires processing both what is sung (lyrics) and how it is performed (musical content).

The independent variability of lyrics and audio creates important ramifications. In most jurisdictions, audio-related aspects (such as melody, harmony, and arrangement) and lyrical composition are protected under separate copyright frameworks, each requiring specialised methods for assessing similarity. Copyright enforcement must distinguish between reproduction of musical elements and plagiarism of lyrical content, a distinction that current litigation often struggles to make reliably. Beyond legal applications, music discovery systems can match songs based on thematic or semantic lyrical connections that transcend acoustic resemblance: a ballad and an uptempo dance track might share profound lyrical similarities about loss and heartbreak that audio-only systems would miss entirely. Creative professionals similarly need tools that identify existing works with similar themes while avoiding unintentional plagiarism, enabling songwriters to understand the landscape of existing content while ensuring originality.

These applications motivate the lyrics matching task: determining whether two musical pieces share similar lyrical content, independent of their acoustic properties. However, evaluating lyrics-matching systems presents a fundamental challenge in MIR research: traditional evaluation requires paired ground-truth lyrics transcriptions, which are scarce due to copyright restrictions and transcription difficulties. This creates a circular problem: without lyrics data, we cannot evaluate lyrics-matching systems, yet we lack the data needed to build and test them. Within this context, MVI, defined as the computational task of automatically determining whether two musical pieces constitute versions of the same composition, serves as an ideal proxy task for several reasons. First, songs that are versions of each other often share lyrical content, providing implicit ground truth labels without requiring explicit lyrics annotations. Second, version identification datasets are more readily available than comprehensive lyrics databases, as they can be constructed from metadata and musical versions databases without needing full lyrical transcriptions. Third, the task naturally encompasses diverse transformation types (musical versions maintaining lyrics, translations preserv-

ing melodies, reinterpretations altering both) that test the robustness of lyrics-matching approaches across realistic scenarios.

By combining lyrics-aware representations with audio content-based representations for version identification, we can develop and evaluate multimodal systems that leverage the complementarity of textual and acoustic signals. This approach addresses both the practical challenge of version identification, in which systems must recognise songs across transformations that affect either lyrics, music, or both, and provides a principled framework for advancing lyrics-matching research despite severe data scarcity.

However, as we explore in the following sections, even within this proxy-task framework, significant challenges remain regarding data availability, methodological limitations, and the lack of standardised evaluation protocols.

Overall, the symbiotic relationship between lyrics matching and version identification makes music an ideal case study for multimodal AI research, exposing fundamental challenges in integrating lyrics and audio that illuminate broader questions about multimodal understanding.

5.1.1 Challenges

Data Scarcity and the Text-Audio Asymmetry — The music domain faces a severe manifestation of the data challenge that characterises multimodal AI development. While vast quantities of musical audio exist in digital form, corresponding textual content — particularly lyrics — is extremely scarce. Copyright restrictions severely limit research access to comprehensive lyrics databases [114]. Manual annotation platforms like Genius[1] cannot scale to cover global music production spanning multiple languages, genres, and cultural contexts. ASR systems, despite recent advances, struggle with the unique challenges of musical content: separating vocals from instrumental accompaniment, handling singing styles that differ significantly from spoken language, processing various vocal effects and production techniques, and dealing with acoustic artefacts introduced by musical production [154]. This asymmetry between abundant audio and scarce text creates a fundamental bottleneck for supervised multimodal approaches that rely on paired data. Even when transcriptions exist, they frequently contain errors, spelling variations, and structural inconsistencies that compromise retrieval quality.

[1] `https://genius.com/`

Methodological Limitations and Lack of Reproducible Baselines — Existing methods for lyrics-based MIR either employ outdated techniques, lack reproducibility, or are evaluated on limited datasets that fail to establish robust benchmarks [147, 246, 59]. The field has been further hindered by the impracticality of relying on pre-existing transcriptions, which are often unavailable, incomplete, or restricted by copyright. A more general challenge involves capturing multilingual semantic relationships, where the same content expressed in different languages should be recognised as similar, while accounting for semantic nuances across linguistic and contextual variations.

Lack of Standardized Evaluation Protocols — While MVI provides a valuable proxy task framework for lyrics-matching research, significant challenges remain in establishing standardised evaluation protocols. Existing version identification datasets vary substantially in size, transformation types, and annotation quality [370], making it difficult to assess whether performance differences reflect genuine methodological advances or dataset characteristics. Furthermore, the proxy task introduces complexities: not all versions maintain lyrical similarity (instrumental versions, radical rewrites), while songs with similar themes but no version relationship create potential false positives for lyrics-based approaches. Without standardised protocols specifying which transformation types to include, how to handle edge cases, and what constitutes fair evaluation, the research community has fragmented into different studies optimising for subtly different objectives.

Moreover, reproducibility remains challenging. Many existing approaches rely on proprietary features, closed datasets, or incompletely documented preprocessing pipelines. Even after code is released, differences in audio quality, segmentation strategies, and feature-extraction implementations can yield substantially different results. The absence of shared reference implementations and standardised preprocessing pipelines means that researchers often cannot reliably reproduce baseline results, making it difficult to assess the true contribution of proposed innovations.

This chapter addresses these challenges by developing WEALY, a fully reproducible end-to-end system for extracting lyrics-aware representations directly from audio using foundation models. By providing complete code, preprocessing pipelines, and evaluation protocols across three diverse datasets, we establish transparent baselines that enable systematic comparison and future research. Our work demonstrates how the strategic use of foundation models can overcome

domain-specific data limitations while validating the complementarity of multimodal fusion, aligning with the broader thesis themes of effective multimodal integration and reproducible research practices.

5.1.2 Our Contribution

This chapter presents several key contributions to the broader thesis investigation of multimodal AI for human expression understanding:

- We develop WEALY (**W**hisper **E**mbeddings for **A**udio-based **LY**rics matching), the first fully reproducible end-to-end pipeline that extracts lyrics-aware representations directly from raw audio without requiring text transcriptions or lyrics databases. By leveraging Whisper's decoder embeddings through contrastive learning on MVI tasks, we demonstrate how strategic use of foundation models can overcome severe domain-specific data limitations.
- We validate the complementarity of lyrics and audio modalities through simple late fusion: combining distance scores from WEALY and audio-content baselines achieves state-of-the-art performance on standard benchmarks (0.912 MAP on SHS100k-v2), substantially outperforming either modality alone. This confirms that lyrics and audio provide distinct information channels that enable more robust version identification when properly integrated.
- We provide fully reproducible implementations, trained models, and transparent evaluation protocols across three diverse datasets, addressing the reproducibility challenges that have limited progress in lyrics-based MIR. Our extensive ablations examine architectural choices, loss functions, and the role of multilingual information in foundation model representations.
- The successful adaptation of Whisper—pretrained on speech recognition—to musical lyrics extraction demonstrates cross-domain transfer capabilities of foundation models, validating a core thesis hypothesis about feature extraction and adaptation across related tasks.

5.1.3 Chapter Structure

The remainder of this chapter proceeds as follows:

- Section 5.2 reviews related work in MVI, lyrics-based retrieval, and multimodal fusion for MIR. We survey audio-based MVI approaches that have achieved strong performance using spectral and harmonic features, examine the limited prior work on lyrics-based matching and its challenges, discuss the role of foundation models like Whisper in music information retrieval, and analyze existing multimodal fusion strategies and their applicability to the lyrics-audio integration problem.
- Section 5.3 describes WEALY in detail, including the problem formulation using MVI as a proxy task for lyrics matching evaluation, the two-stage pipeline design comprising Whisper-based feature extraction and transformer-based representation learning with contrastive objectives, experimental setup across three datasets with comprehensive baselines, results demonstrating superiority over transcription-based methods, ablation studies examining architectural choices and design decisions, and multimodal fusion experiments showing complementarity between lyrics and audio modalities.
- Section 5.4 concludes with a summary of contributions and future research directions, including extensions to other music information retrieval tasks, integration with explainability methods to understand what lyrical and/or audio features drive similarity judgments, and applications beyond music to other domains involving sung or chanted speech.

5.2 Related Work

5.2.1 Musical Version Identification

MVI aims to determine whether two musical recordings represent versions of the same underlying composition. The task has motivated significant research in MIR due to applications in copyright detection, music cataloging, and discovery systems [370]. Audio-based MVI systems have achieved strong retrieval performance using learned audio embeddings. The ByteCover series [76, 78] made important advances through multi-sequence similarity learning to handle variable-length audio inputs. ByteCover2 introduced dimensionality reduction of latent embeddings for improved efficiency, while ByteCover3 optimized for short query scenarios, with ByteCover1/2 achieving 0.813±0.006 MAP and ByteCover3.5 reaching 0.857 MAP on SHS100k-v2. The CLEWS framework [294] represents the current state-of-the-art in audio-content-based MVI. CLEWS advances beyond previous approaches through contrastive learning from weakly-labeled audio segments. Unlike methods requiring precise alignment, CLEWS operates on variable-length sequences with only version-level labels, learning to identify relevant matching segments automatically through a specialized loss function that encourages matching segments to have high similarity while ensuring non-matching segments from different songs remain distant. CLEWS achieves 0.876±0.005 MAP on SHS100k-v2, establishing the strongest baseline for audio-content approaches. However, audio-content approaches face limitations when versions diverge significantly in musical arrangement while maintaining lyrical similarity, or when instrumental versions eliminate vocal cues entirely. These limitations motivate investigation of lyrics-based and multimodal approaches that can capture complementary information.

5.2.2 Lyrics-Based Music Retrieval

Lyrics-based approaches to musical similarity and version identification have received limited attention, primarily due to data scarcity challenges. The existing literature can be categorized into text-based methods requiring pre-existing transcriptions and audio-based methods that extract lyrical information directly.

Early text-based methods treated lyrics matching as an information retrieval problem. Knees et al. [147] explored lyrics alignment for automatic retrieval from web sources, revealing critical data quality issues: roughly 43% of web-scraped lyrics were irrelevant, with frequent spelling variations, misheard lines, and struc-

tural inconsistencies. Patra et al. [246] developed a TF-IDF-based recommendation system for Hindi songs, but required manual collection of over 31,000 lyrics with extensive preprocessing to handle romanization issues, demonstrating the labor-intensive nature of curating quality lyrics datasets.

These text-based approaches face fundamental limitations: dependence on pre-existing transcriptions limits coverage, transcription quality varies substantially, manual curation is impractical at scale, and simple string matching struggles to capture semantic similarity across paraphrases or translations.

The integration of singing voice recognition represents an intermediate approach. Vaglio et al. [339] combined automatic singing voice recognition with string similarity measures for musical versions detection, but singing voice recognition proved substantially more challenging than speech recognition due to melodic variations, sustained vowels, and harmonic interference. Correya et al. [59] proposed a multi-stage pipeline integrating metadata, TF-IDF lyrics from Musixmatch, and audio features, demonstrating that lyrics could complement audio-based methods but encountering limited coverage and transcription quality issues.

Recent work leveraging modern ASR has begun addressing transcription quality limitations. Zhuo et al. [385] demonstrated that Whisper [265], combined with large language models for post-processing, achieves transcription quality competitive with human annotations. Balluff et al. [16] created the LyricCovers 2.0 dataset by applying Whisper to 78,862 tracks spanning 80 languages, exploring lyrics-based version identification using Jina embeddings trained with triplet loss. While demonstrating modern ASR viability, performance remained substantially below audio-content baselines [16], and multi-stage pipelines (source separation, ASR transcription, embedding extraction) introduced complexity and error accumulation.

Du et al. [77] recently presented X-Cover, but their work lacks transparency regarding architectural details, training procedures, and evaluation protocols, limiting reproducibility and scientific value.

These approaches demonstrate lyrics information's potential value but face persistent challenges: transcription dependency, singing voice recognition difficulties, multi-stage pipeline complexity, and limited reproducibility. These limitations motivate our end-to-end approach that extracts semantic lyric representations directly from audio while maintaining full transparency.

5.2.3 Speech Foundation Models for Music Information Retrieval

The emergence of large-scale foundation models pretrained on diverse audio and speech data has opened new possibilities for extracting meaningful representations without task-specific annotations, potentially circumventing the data scarcity issues that plague supervised approaches.

Whisper [265] represents a particularly relevant foundation model for lyrics-based retrieval. Trained on 680,000 hours of multilingual and multitask supervised data, Whisper employs a transformer encoder-decoder architecture where the encoder processes mel-spectrograms to produce acoustic features, while the decoder generates text transcriptions autoregressively. Whisper's decoder produces hidden states at each generation step that, when applied to music, represent the model's internal understanding of lyrical content before committing to specific word predictions. These hidden states encode semantic content, phonetic patterns, and linguistic structure in continuous embeddings that can capture nuanced similarities, unlike final token predictions which collapse this rich representation into discrete words.

Several works have begun exploring Whisper for music information retrieval. Zhuo et al. [385] demonstrated high-quality lyrics transcription combining Whisper with Large Language Model (LLM)-based post-processing. Balluff et al. [16] used Whisper transcriptions as input to sentence transformers for version identification. Du et al. [77] reported (with limited detail) that Whisper features could improve MVI performance. However, systematic investigation of how to optimally extract and adapt Whisper representations for retrieval tasks remains limited.

Our work addresses critical open questions: which layers and components provide the most informative representations for semantic similarity, how to aggregate variable-length sequences into fixed-size representations, what training objectives best adapt Whisper embeddings to retrieval tasks, and whether Whisper's multilingual training enables cross-lingual version identification. We provide systematic answers through comprehensive ablation studies and rigorous evaluation across multiple datasets.

5.2.4 Multimodal Fusion in Music Information Retrieval

Multimodal approaches that combine audio content and lyrics information remain underexplored in music information retrieval research. The fundamental challenge involves developing fusion strategies that effectively integrate complementary in-

formation: a musical version song might preserve lyrics while dramatically altering musical style, maintain melody while translating lyrics, or simultaneously transform both dimensions while retaining recognizability.

Most existing fusion work in MIR focuses on combining acoustic features with metadata (artist, genre, year) or social tags rather than lyrics [59]. Recent deep learning approaches have begun exploring attention mechanisms for cross-modal fusion [357], but application to lyrics-audio integration for version identification remains limited, with most work focusing on other tasks like music emotion recognition [384].

For MVI, an additional challenge arises from asymmetric modality performance: current audio-content features [294, 77] achieve substantially stronger unimodal performance than lyrics-based features on standard benchmarks [16, 77], suggesting that naive fusion may simply defer to the stronger audio modality rather than leveraging complementary lyrical information.

Our work addresses these gaps by systematically evaluating fusion strategies for lyrics-audio integration, quantifying modality complementarity, and demonstrating whether lyrics provide information genuinely orthogonal to audio content in the context of version identification.

Summary and Positioning

Existing work in MVI has achieved strong results using audio-content approaches [76, 78, 294], but faces limitations when musical transformations preserve lyrics while altering audio significantly. Lyrics-based methods [147, 246, 59, 339] provide a complementary retrieval signal but have been hindered by data scarcity, transcription errors, and lack of reproducible baselines. Recent foundation models like Whisper [265] create opportunities for extracting lyrics-aware representations directly from audio [385, 16, 77], yet systematic investigation of optimal extraction and adaptation strategies remains absent. Finally, multimodal fusion strategies that effectively combine lyrics and audio information have received minimal attention despite their potential to improve robustness. Our work addresses these gaps by: (1) developing a fully reproducible pipeline (WEALY) for extracting lyrics-aware representations from audio using Whisper decoder embeddings; (2) establishing rigorous baselines through systematic evaluation across three diverse datasets (DiscogsVI-YT, SHS100k-v2, LyricCovers2.0); (3) conducting comprehensive ablation studies to understand the impact of archi-

tectural choices, loss functions, pooling strategies, and multilingual information; and (4) demonstrating through simple multimodal fusion that lyrics and audio provide strongly complementary information for version identification, validating the multimodal integration hypothesis central to this thesis.

5.3 Leveraging Whisper Embeddings for Audio-based Lyrics Matching

This section presents our work on audio-based lyrics matching through Whisper decoder embeddings, conducted in collaboration with Joan Serrà, Paolo Torroni, and Yuki Mitsufuji [200].

Key Contribution: We introduce WEALY (**W**hisper **E**mbeddings for **A**udio-based **LY**rics matching), the first reproducible end-to-end pipeline that extracts lyrics-aware representations directly from raw audio without requiring intermediate text transcription.[a] Operating under the assumption that musical versions sharing lyrical content provide a lower bound on detectable semantic similarities, we employ MVI as a proxy task for evaluating lyrics matching capabilities. The work addresses a critical gap in MIR by establishing a transparent and scalable approach to lyrics-based matching that consistently outperforms transcription-based methods. Through extensive experiments on three large-scale datasets—DiscogsVI-YT (DVI), SHS100k-v2 (SHS), and LyricCovers2.0 (LYC)—we demonstrate that learned representations from Whisper's decoder capture meaningful lyrical semantics that benefit from the model's inherent multilingual capabilities.

[a] Code and pre-trained checkpoints are available at `https://github.com/helemanc/audio-based-lyrics-matching`

We further show that combining WEALY with audio content-based methods through simple late fusion yields substantial improvements on MVI, highlighting the complementarity of lyric and audio signals and suggesting promising directions for multimodal MIR systems.

5.3.1 Method

Problem Formulation

We formalise audio-based lyrics matching as the task of identifying semantic and structural similarities between songs based on their lyrical content, where songs are considered similar if they share thematic connections, semantic overlaps, or structural patterns in their lyrics. Given the absence of dedicated datasets for direct lyrics matching evaluation, we employ MVI as a proxy task, operating

under the assumption that musical versions sharing lyrical content should provide a lower bound on detectable semantic similarities.

Pipeline

WEALY employs a two-stage pipeline design for end-to-end lyrics matching, based on feature extraction and adaptation. For feature extraction, we process raw audio through our Whisper decoder embeddings extraction pipeline, producing dense tensor representations that capture the semantic content of song lyrics without requiring intermediate text transcription. For feature adaptation, we input lyrical representations into a transformer-based architecture trained with contrastive learning on the MVI task.

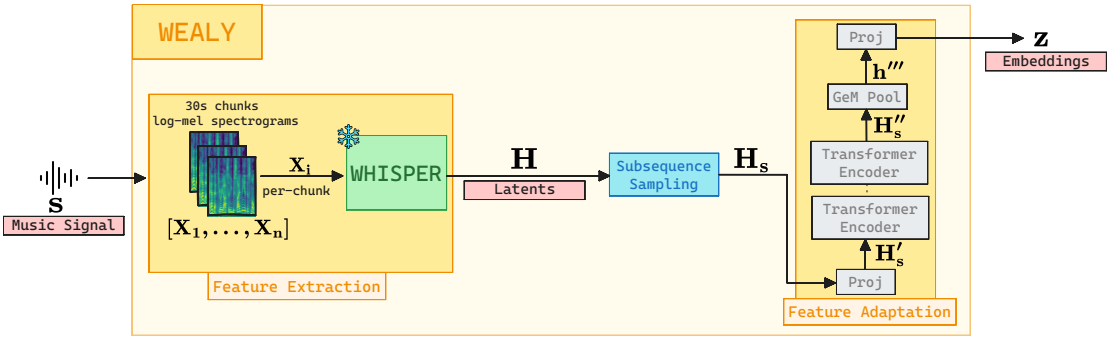


Figure 5.1: *Overview of the WEALY architecture. Stage 1 extracts lyrical latents with Whisper, while stage 2 learns contextualized representations with a transformer encoder and projects them into a compact embedding space.*

Feature Extraction — Our approach begins with audio preprocessing, converting to mono, resampling files to 16 kHz, and truncating to a maximum of 5 min to balance computational efficiency with content coverage. Unlike some prior work that applies vocal source separation before feature extraction, we operate directly on the raw mixture. Since some results suggest that even with a strong separator Whisper's transcription quality only improves modestly [16], we avoid source separation in our pipeline.

Whisper employs a classic transformer architecture with stacked encoder and decoder blocks, where attention mechanisms propagate information between components as the system processes an input audio file $\mathbf{s}$, splitting it into n 30-second overlapping chunks so that $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]$. From each chunk, Whisper extracts log-mel spectrograms $\mathbf{X}_i \in \mathbb{R}^{t \times 128}$, yielding a sequence of spectrograms $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$, where t depends on the frame length (Fig. 5.1). For each

log-mel spectrogram $\mathbf{X}_i$, the encoder processes the raw audio and captures positional information for spoken elements while preserving temporal structure. The decoder then leverages this encoded information to predict tokens (words) in an autoregressive manner, using both audio representations and previously predicted tokens to build contextual understanding throughout the sequence.

From each $\mathbf{X}_i$, we extract decoder representations $\mathbf{H}_i = [\mathbf{h}_i^{(j)}, \dots \mathbf{h}_i^{(m_i)}]$, where m_i is the number of per-chunk hidden decoder states or latents. Since Whisper processes one chunk at a time, we concatenate the hidden representations after each autoregressive pass over one chunk to obtain the final representation. After autoregressive decoding, we concatenate all latents from all chunks into a single matrix $\mathbf{H} = [\mathbf{h}_1^{(1)}, \dots \mathbf{h}_1^{(m_1)}, \mathbf{h}_2^{(1)}, \dots \mathbf{h}_n^{(m_n)}]$. This results in $\mathbf{H} \in \mathbb{R}^{m \times d_w}$, where $m = \sum_{i=1}^{n} m_i$ is the total number of latents across all chunks and $d_w = 1280$ is the latent dimension of Whisper-turbo's 4-layer decoder. The length m varies because Whisper produces variable-length sequences of hidden states, skipping over silent regions and generating outputs only for active segments. Thus, the final number of hidden states depends on the audio content characteristics rather than being determined solely by the track duration or a fixed segmentation strategy.

This autoregressive process enables each latent $\mathbf{h}$ to incorporate prior context, thereby supporting coherent transcriptions that accurately reflect linguistic dependencies and semantic relationships. Context resets are triggered by temperature thresholds during uncertain generation, creating distinct segments in the hidden-state sequence. We extract representations from Whisper's final decoder layer before token sampling, capturing the model's refined semantic understanding of lyrical content: the compressed knowledge about the song's lyrics before committing to specific word predictions. What we obtain can be cast as 'lyrics-aware Whisper latents', which we use as representations of the music tracks. Our implementation selects hidden states corresponding to the highest-scoring beam search sequence through modifications to the original Whisper codebase. Additionally, based on the results obtained by [385], we investigate initialisation prompting effects by comparing standard tokens against task-specific prompts such as *"lyrics:"*. Since our results aligned with their findings showing improved representation quality with task-specific prompts, we consistently employed the *"lyrics:"* prompt in all subsequent experiments (Sec. 5.3.3).

Feature Adaptation — The second stage of WEALY employs a transformer-based architecture to learn similarity representations from the Whisper-derived

lyrical latents (Fig. 5.1). Given the feature sequence $\mathbf{H}$ extracted in the first stage, we sample random subsequences $\mathbf{H}_s \in \mathbb{R}^{k \times d_w}$ of fixed length $k = 1500$ to ensure computational efficiency and to expose the model to diverse temporal segments of each track. This subsequence length was chosen empirically: shorter windows (e.g., $k = 500$ or 1000) failed to capture sufficient lyrical context for semantic alignment, while longer windows ($k = 2000$) degraded performance.

Each subsequence $\mathbf{H}_s$ is first projected into the model dimension d_h through a linear layer, yielding $\mathbf{H}'_s \in \mathbb{R}^{k \times d_h}$. The sequence is then processed by a stack of transformer encoder blocks, which produce contextualized representations $\mathbf{H}''_s \in \mathbb{R}^{k \times d_h}$ without altering the sequence length. To obtain a temporally-collapsed vector suitable for generic similarity computation, we apply Generalized Mean (GeM) pooling [263] across the temporal axis, resulting in a single representation $\mathbf{h}''' \in \mathbb{R}^{d_h}$. Finally, a linear projection maps $\mathbf{h}'''$ into the target embedding space, yielding a compact semantic representation $\mathbf{z} \in \mathbb{R}^{d_e}$. Our final configuration comprises $N = 4$ transformer encoder blocks. Each block uses a self-attention model dimension of $d_h = 768$ with 12 attention heads and a feed-forward MLP dimension of 1024. A dropout $p = 0.1$ is applied. Input subsequences are first linearly projected to d_{model}, processed by the N blocks, GeM-pooled over time, and finally projected to an embedding of size $d_e = 512$.

To learn our representation $\mathbf{z}$, we adopt a contrastive learning paradigm based on the NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss [50]. Let $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ denote the cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$. For a positive pair of songs i and j, the loss is defined as

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)},$$

where $\tau = 0.1$ is a temperature parameter and $\mathbf{1}_{[k \neq i]}$ is an indicator function that excludes self-similarity. The final loss is computed symmetrically across all positive pairs (i, j) and (j, i) in the mini-batch. This encourages embeddings of versions of the same song to cluster together while pushing apart those from different songs.

Ablation Studies

The architecture in Fig. 5.1 represents the final design we converged upon after extensive experimentation with different pooling strategies, projection schemes,

transformer configurations, and loss functions. In our results we report on ablation studies across key design choices: loss function, pooling strategy, and multimodal fusion.

Loss function — We compare NT-Xent against the triplet loss [291], which uses margin-based ranking between positive and negative pairs, and the CLEWS loss [294], originally designed for multi-sequence similarity learning.

Pooling strategy — We evaluate simple average pooling against GeM Pooling, which applies learnable power-mean pooling to emphasise informative temporal regions. We also explore using the canonical CLS token representation instead of pooling across timesteps and directly averaging Whisper embeddings and passing them through an MLP before contrastive training (bypassing the transformer encoder).

Language-constrained representations — To assess the role of multilingual cues in Whisper, we force the model to transcribe all songs into English tokens exclusively before generating latents. This setting allows us to test whether removing cross-lingual information degrades retrieval performance and to quantify the benefit of retaining multilingual features in the embeddings.

Multimodal fusion — We also conduct a small multimodal fusion experiment, where WEALY (lyrics-aware) and CLEWS (audio-content) are combined via late fusion at the distance level. For each song pair (i, j), we compute distances $\delta_{\text{CLEWS}}(\mathbf{s}^{(i)}, \mathbf{s}^{(j)})$ and $\delta_{\text{WEALY}}(\mathbf{s}^{(i)}, \mathbf{s}^{(j)})$, and define the combined distance as $\delta(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) = \delta_{\text{CLEWS}}(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) + \alpha \cdot \delta_{\text{WEALY}}(\mathbf{s}^{(i)}, \mathbf{s}^{(j)})$. We use the original bpwr formulation for δ_{CLEWS} [294], the cosine distance for δ_{WEALY}, and we set $\alpha = 1.5$ based on validation experiments.

5.3.2 Experimental Setup

Datasets

We evaluate WEALY on three publicly-available datasets: DiscogsVI-YT (DVI) [9], SHS100k-v2 (SHS) [377], and LyricCovers2.0 (LYC) [16]. All audio is processed at 16 kHz mono with 5-minute length constraints. SHS represents a well-established reference dataset, although it exhibits an important bias towards unrealistically large version group sizes [294] and YouTube link dependencies limit the collection of all entries (we were able to collect 82% of it). DVI addresses these limitations by providing a more recent dataset that is five times

larger and with more realistic version group size distributions, offering improved validity. LYC encompasses 78,862 tracks (54,301 covers and 24,561 originals) with annotated lyrics via `genius.com` links. In its original form, the dataset contains duplicated entries within version groups, since tracks are paired with both Whisper-generated and annotated transcriptions where available. This design reflects one of the original authors' goals, namely to evaluate transcription accuracy and to compare the impact of different transcription sources on MVI. In our case, since the focus is on evaluating audio-derived lyric representations rather than transcription quality, we reduced the dataset to unique clique−version pairs while retaining only the metadata needed for retrieval. The dataset spans 80 languages (82.6% English) with 95% of musical versions maintaining the original language, making it relevant for validating multilingual lyrics-matching aspects. We note that while Whisper's training corpus is undisclosed, potential overlap with our evaluation data does not constitute contamination, as Whisper serves only as a frozen feature extractor and our train/test splits apply exclusively to WEALY's contrastive learning stage. The task we evaluate (version identification) differs fundamentally from Whisper's transcription objective, and any exposure to these recordings during Whisper's pre-training would enhance rather than confound feature quality.

Baselines

We consider an extensive set of baselines covering different approaches to lyrics-based matching. TF-IDF baselines follow Correya et al. [59], comparing transcriptions either directly with cosine similarity or with a Lucene-style "More Like This" query (we denote these baselines as TF-IDF-Cosine and TF-IDF-Lucene, respectively). Another baseline encodes Whisper ASR transcriptions using Sentence-BERT (SBERT) and compares them using both cosine similarity and a learnt Transformer using NT-Xent (we denote the two baselines as ASR-SBERT-Cosine and ASR-SBERT-Trasf, respectively). We also include a representation-only baseline where we directly average Whisper decoder embeddings across time and compute similarities with cosine distance, without any further training (denoted as Whisper-AvgEmb). This baseline isolates the quality of raw representations extracted from Whisper before our proposed adaptation. Finally, we consider a random baseline that assigns distances uniformly at random, and a Non-instrumental Oracle that aims to account for instrumental versions that are not detectable

through a lyrics-based approach. This oracle provides an upper bound by setting distances to zero only for same-clique songs with 'valid' transcriptions, where validity is enforced through a dedicated process that cleans text and filters out unreliable cases associated to instrumental versions. For a song to be considered a match, the oracle requires it to have (i) a minimum length of 10 words, (ii) at least five alphanumeric characters, (iii) limited bigram/trigram repetition (no more than 70% repetition with at least three unique bigrams and two unique trigrams), (iv) no single phrase repeated in more than half of the text, and (v) no purely musical content such as annotations, humming, or repetitive syllable patterns.

Training and Evaluation

We train all models using the AdamW optimizer with a learning rate of 10^{-4}, a weight decay of 10^{-3}, a cosine schedule with 50 warmup epochs and a minimum learning rate of 10^{-6}. Training is carried out for a maximum of 1000 epochs with a batch size of 64 across four GPUs. To prevent overfitting, we apply early stopping with a patience of 20 epochs, monitoring mean average precision (MAP) on the validation set as the stopping criterion. We also use MAP to report results on the test set, which conforms to standard practice [370].

As mentioned, during training, we sample random subsequences of length $k = 1500$ tokens from the Whisper representations $\mathbf{H}$ to improve robustness and provide diverse temporal coverage of each track. However, at validation time, we adopt a deterministic procedure, always selecting the first k tokens from each track to ensure consistency across epochs. At test time, we extract overlapping subsequences of k tokens with 90% overlap from both query and candidate tracks. Similarities for each query and candidate pair are computed on a full-track basis, using the two corresponding chunk-based embedding sequences and taking the maximum pairwise cosine similarity. This best-match approach simulates a chunk-based retrieval system that exploits the most discriminative portions of each track.

5.3.3 Results

Table 5.1 first highlights the performance of transcription-based baselines (TF-IDF and ASR-SBERT variants), which provide limited retrieval quality. The representation-only baseline (Whisper-AvgEmb) performs even worse, showing

that averaged decoder embeddings do not generalise, and that a dedicated model is needed to learn effective representations. Overall, WEALY consistently outperforms all transcription-based methods, establishing a reproducible lyrics-aware baseline that scales with dataset size. For completeness, we here also note the results reported by [77], which introduces two Whisper-based variants: "Whisper-AR", described as an autoregressive feature extraction approach, and "Whisper-E", obtained by fine-tuning with adapters. The terminology in [77] is somewhat ambiguous, but the reported MAP on SHS is 0.708 for Whisper-AR and 0.437 for Whisper-E. While a direct comparison is difficult due to the unclear methodology, our results are broadly aligned with these numbers, with the difference that WEALY offers a transparent and fully reproducible end-to-end pipeline.

Method	**DVI**	**SHS**	**LYC**
Random	0.001 ± 0.000	0.003 ± 0.003	0.002 ± 0.002
Non-instrumental Oracle	0.967 ± 0.000	0.956 ± 0.004	0.954 ± 0.004
TF-IDF-Cosine	0.272 ± 0.002	0.503 ± 0.008	0.537 ± 0.009
TF-IDF-Lucene	0.242 ± 0.002	0.457 ± 0.008	0.486 ± 0.009
ASR-SBERT-Cosine	0.294 ± 0.002	0.508 ± 0.008	0.573 ± 0.009
ASR-SBERT-Trasf	N/A	0.480 ± 0.001	0.516 ± 0.008
Whisper-AvgEmb	0.166 ± 0.001	0.297 ± 0.007	0.322 ± 0.007
WEALY	**0.328 ± 0.002**	**0.640 ± 0.008**	**0.692 ± 0.008**

Table 5.1: *Main results: MAP on the three considered datasets.*

Table 5.2 reports the ablation study using SHS. First, the choice of loss function is critical: NT-Xent clearly outperforms both triplet loss and CLEWS loss, the latter being originally designed for multi-sequence similarity but proving less effective in this setting. Regarding pooling and architecture, simple averaging and the CLS token are competitive, but consistently below the GeM pooling baseline. In contrast, directly averaging Whisper embeddings and passing them through an MLP performs substantially worse, confirming that temporal modelling is essential for capturing lyric semantics. The language-constrained experiment further shows that multilinguality is a strength of Whisper: restricting decoding to English leads to a noticeable performance drop, indicating that multilingual cues in the latents provide useful information for retrieval.

Finally, Table 5.3 focuses more on MVI and demonstrates that multimodality brings clear improvements. A simple late-fusion of WEALY and CLEWS at the distance level yields a performance that surpasses both unimodal approaches, indicating that the two models capture complementary information. This improvement is obtained without introducing additional model complexity, relying

Category	Method	MAP
Default	**WEALY**	**0.640 ± 0.008**
Loss functions	WEALY - Triplet loss	0.548 ± 0.008
	WEALY - CLEWS loss	0.450 ± 0.008
Pooling strategies	WEALY - Simple average	0.627 ± 0.008
	WEALY - CLS token	0.621 ± 0.008
	WEALY - Average+MLP	0.389 ± 0.008
Language settings	WEALY - English only	0.578 ± 0.008

Table 5.2: *Ablation study of WEALY on the SHS dataset.*

only on a straightforward distance-level combination. We note that the optimal fusion coefficient ($\alpha = 1.5$) assigns higher weight to WEALY despite its lower individual performance (MAP=0.640) compared to CLEWS (MAP=0.876). This coefficient was determined empirically from a finite grid search on the validation set, and while the asymmetry is surprising, a deeper analysis of the fusion dynamics is left to future work. Overall, while multimodal fusion and MVI are not the primary focus of this work, these results highlight the potential of lyrics matching for advancing the latter.

Method	MAP
ByteCover1/2 (as reported by [294])	0.813 ± 0.006
ByteCover3.5 [77]	0.857
CLEWS [294]	0.876 ± 0.005
WEALY+CLEWS	**0.912 ± 0.004**

Table 5.3: *Comparison of audio content-based MVI approaches with the proposed multimodal approach the SHS dataset.*

5.3.4 Findings

We introduced WEALY, the first reproducible end-to-end pipeline for audio-based lyrics matching. WEALY builds on Whisper decoder embeddings, which are extracted directly from raw audio without relying on intermediate transcriptions, and leverages the ground truth of MVI datasets to train a transformer encoder via contrastive learning. This design allows us to capture lyrics-aware representations that are robust, scalable, and directly optimised for retrieval. Our experiments show that WEALY consistently outperforms transcription-based pipelines, scales effectively to large datasets, and benefits from the latent multilingual properties of Whisper, which preserves useful cross-lingual cues for matching. While results

remain below the strongest audio content-based models for MVI, we observe that fusing WEALY with them through a simple distance-level combination can lead to clear improvements, underlining the complementarity of lyric and audio cues, and pointing to promising multimodal extensions for MVI. Overall, our findings establish WEALY as a strong and transparent baseline for lyrics matching, with direct implications for MIR tasks such as version identification, copyright detection, and music discovery.

5.4 Conclusion

This chapter addresses the fundamental challenge of audio-based lyrics matching by introducing WEALY, the first fully reproducible end-to-end pipeline that extracts lyrics-aware representations directly from raw audio without requiring pre-existing transcriptions. By leveraging Whisper's decoder embeddings and adapting them through contrastive learning, our approach circumvents the severe lyrics scarcity problem that has hindered prior work in MIR. Through systematic evaluation across three diverse datasets, we establish that WEALY consistently outperforms transcription-based methods, and comprehensive ablation studies reveal optimal design choices including NT-Xent loss, GeM pooling, and preservation of multilingual capabilities. Most significantly, simple late fusion of WEALY with audio-content baselines achieves 0.912 MAP, representing a +4.1% absolute improvement over the stronger unimodal baseline, validating that lyrics and audio provide orthogonal information channels that enable a more robust version identification when properly integrated.

This work complements and extends the broader thesis investigation of multimodal AI for human expression in several critical ways. First, it further demonstrates that the severe domain-specific data limitations of the multimodal domains we address (political, clinical, music) can be overcome by employing pretrained foundation models. Second, the music domain provides the clearest validation of multimodal complementarity: unlike political discourse, where speech and text represent parallel channels or clinical settings where modalities may correlate, music presents truly independent dimensions where lyrics and audio can vary freely, and the substantial performance gains from simple fusion demonstrate that when modalities genuinely capture orthogonal information, even naive fusion strategies yield robust improvements. Third, our late fusion approach provides a degree of interpretability by exposing separate distance scores, allowing analysts to examine whether similarity stems from lyrics, audio, or both—a useful capability for copyright contexts where different legal frameworks protect textual and musical content. However, this represents only a first step toward explainability: understanding modality contributions at the aggregate level does not reveal which specific lyrical phrases, melodic patterns, or temporal segments drive similarity judgments. Therefore, substantial technical advances are necessary before multimodal systems can provide the granular explanations required for clinical trust or

legal proceedings. Fourth, the comprehensive ablation studies and reproducible baselines established here constitute a rigorous methodological blueprint applicable across domains, providing a template that addresses the fragmentation limiting progress in multimodal research.

These contributions validate core thesis hypotheses. However, our analysis also suffers from several limitations. WEALY's performance below audio-content baselines indicates room for exploring different foundation models; our experiments employed only simple late fusion while more sophisticated strategies warrant investigation, and evaluation focused on version identification as a proxy task rather than dedicated lyrics similarity datasets. Nevertheless, the demonstration that foundation model adaptation, simple fusion strategies, and systematic evaluation can yield substantial improvements in this challenging domain reinforces the central thesis argument: understanding human expression requires multimodal approaches that integrate complementary information sources while maintaining transparency about their contributions.

Link to Research Questions. This chapter primarily addresses **RQ9** and **RQ10**, while also contributing to the cross-domain **RQ1** and **RQ3**. Together, these findings establish the methodological and empirical foundations for multimodal music understanding through lyrics-aware audio representations.

- **RQ1: Multimodal Necessity** — The results confirm that lyrics and audio provide orthogonal yet complementary information channels, demonstrating that multimodal integration captures aspects of musical similarity that unimodal systems overlook.
- **RQ3: Systematic Solutions** — The introduction of the WEALY pipeline and the accompanying ablation studies offer a reproducible methodological framework applicable to multimodal learning beyond music, addressing the fragmentation that often limits cross-domain progress.
- **RQ9: Large-Scale Lyrics Data Challenge** — Addressed through the use of Whisper-based decoder embeddings and contrastive adaptation, which eliminate the dependence on transcriptions and overcome data scarcity constraints in large-scale lyrics analysis.
- **RQ10: Multimodal Performance** — The simple late fusion of WEALY with audio-content baselines achieves substantial improvements in version identification accuracy, validating the benefits of integrating lyrics-aware and audio-based representations for music information retrieval.

CHAPTER 6

Roadmap Towards Explainable Multimodal AI

6.1 Introduction

A consistent need for explainability in multimodal systems incorporating speech has emerged across the three application domains explored in the preceding chapters.

In multimodal argument mining (Chapter 3), explainability must reveal how linguistic and paralinguistic features combine to drive predictions, not just their individual contributions. This enables both theoretical validation of multimodal persuasion frameworks and practical tools for citizens to assess whether political influence stems from logical content, rhetorical delivery, or their strategic interaction.

In clinical decision support (Chapter 4), practitioners need to understand not just whether linguistic markers or acoustic features dominate, but how these modalities interact to produce diagnostic predictions that guide treatment decisions.

In musical version identification (Chapter 5), copyright disputes require dis-

tinguishing whether similarity derives from lyrical content, audio characteristics, or their combination, as these correspond to different legal frameworks and ownership structures.

This convergent demand motivates systematic investigation of XAI for speech-based multimodal systems.

We argue that deriving useful explanations for speech and audio requires working at the level of human auditory perception rather than at the level of technical feature representations. This need stems from a dual asymmetry: it concerns both how features are represented across modalities and how humans process and verbalise explanations about them. Features that are easily interpretable in text or vision often lack a direct cognitive counterpart in audio. In speech and sound, the features that models manipulate — such as spectral coefficients, energy contours, and temporal envelopes — do not align with how humans naturally perceive and conceptualise auditory experience, making interpretability inherently more challenging.

This asymmetry becomes evident when we examine how humans process explanations across different modalities: our ability to not only perceive but also describe and explain what we perceive varies dramatically. In text classification, highlighting relevant words produces explanations we can both read and articulate. For instance, we explain that a document is argumentative because it contains phrases like "therefore," which signal logical connections. In image classification, saliency maps highlight regions we can both see and describe with precision. We identify "a red octagonal shape with white letters spelling 'STOP'" and explain why this indicates traffic control. Both modalities offer a critical advantage: we possess rich descriptive vocabularies that transform perception into explanation, enabling us to articulate not just what we perceive but why it matters.

Audio differs fundamentally. We perceive temporal patterns, spectral characteristics, intensity variations, rhythmic structures, and textural qualities through direct auditory experience. Yet translating these perceptions into precise language proves remarkably difficult. When explaining what makes speech sound "sad" versus "energetic," we resort to vague terms like "flat," "monotone," or "breathy," lacking the vocabulary for formant frequencies, spectral slopes, or jitter that actually shape our experience. This cognitive limitation means that standard explanation methods fail: a saliency map highlighting spectrogram regions tells clinicians little about depressive speech patterns they can validate, and attention weights provide

argumentation scholars no insight they can reconcile with rhetorical devices.

These intertwined representational and cognitive asymmetries call for a different paradigm of explainability in speech-based systems.

We identify *listenability* as a critical property: explanations should be audio signals that domain experts can listen to and evaluate through native auditory perception. But what makes an audio explanation truly listenable? We address five fundamental questions:

- **Temporal localisation:** Can the explanation identify precisely when the classifier attends to diagnostic features?
- **Audio quality:** Does it preserve acoustic fidelity without perceptual artefacts?
- **Faithfulness:** Does it accurately reflect which audio regions influence decisions?
- **Discriminability:** Can listeners distinguish between explanations for different classes?
- **Actionability:** Can domain experts use what they hear to validate or improve the system?

Through experimental work on environmental sound classification, we demonstrate that high-quality listenable explanations require addressing phase estimation. Current methods for deriving listenable explanations operate on magnitude spectrograms, which lack the phase information needed for audio reconstruction. Converting these magnitude-only representations back to audio requires phase estimation, a process that introduces perceptual artefacts, degrading explanations' audio quality. By incorporating speech separation architectures that work directly in the time domain and preserve phase coherence throughout, we avoid the reconstruction phase entirely, producing explanations that are temporally localised, artefact-free, and faithful to classifier decisions.

This transforms explainable AI from a technical debugging tool into a practical instrument: clinicians can hear isolated acoustic characteristics like flattened prosody or reduced speech rate and validate them through direct auditory experience, while argumentation scholars can aurally identify paralinguistic cues like emphatic stress or rhetorical pauses as persuasive devices.

Yet fundamental questions remain that define the roadmap ahead: do listenable explanations truly enable actionable understanding in high-stakes contexts? Can perceptually accessible explanations overcome the inherent difficulty humans face in describing what they hear, or does this cognitive limitation ultimately constrain their utility for clinical validation and legal proceedings?

This chapter establishes a roadmap towards explainable multimodal AI by defining listenability as a fundamental property that audio explanations must possess, just as text explanations must be readable and image explanations must be visible. Without this foundational principle, progress towards interpretable speech-based systems cannot proceed meaningfully. We validate this principle through proof of concept on environmental sound classification, demonstrating that perceptually accessible explanations can maintain technical faithfulness while enabling auditory validation. However, realising the full roadmap remains future work: extending listenable explanations to speech-based multimodal systems across political discourse, clinical assessment, and music; combining audio and text explanations for holistic multimodal understanding; and demonstrating utility for domain experts in high-stakes deployment contexts. By establishing listenability as both measurable and achievable, this work provides the necessary foundation for advancing explainable multimodal AI beyond this initial proof of concept.

6.1.1 Our Contribution

This chapter presents our work on generating high-quality *listenable explanations* for audio classifiers. We propose LMAC-TD, a post-hoc explanation method that produces auditory interpretations directly in the time domain, allowing experts to *hear* which acoustic properties drive model decisions. Building on the L-MAC framework [242], our approach improves the quality and faithfulness of listenable explanations by generating interpretable audio signals that remain perceptually aligned with the classifier's reasoning. We validate the method on environmental sound classification tasks and show that it yields explanations that are both faithful to model predictions and perceptually meaningful to human listeners. Overall, this work establishes *listenability* as a measurable and attainable property of audio explanations, marking a step toward perceptually grounded interpretability. The framework also opens the way to extending listenable explanations to speech-based multimodal systems explored in this thesis, where perceptually grounded interpretability is essential for understanding and trust.

6.1.2 Chapter Structure

The remainder of this chapter proceeds as follows:

- Section 6.2 reviews related work in XAI, with a focus on audio-based interpretability. It outlines intrinsic and post-hoc methods, discusses concept- and prototype-based approaches, and highlights the limitations of current explanation techniques.
- Section 6.3 presents the LMAC-TD methodology, describing its architecture, training objective, and experimental setup. We outline how generating explanations directly in the time domain overcomes the limitations of frequency-based approaches and report quantitative and perceptual evaluations demonstrating the effectiveness of the proposed method.
- Section 6.4 concludes with a discussion of implications for explainable multimodal AI and outlines future directions toward perceptually grounded explanations for speech-based systems.

6.2 Related Work

XAI has emerged as a critical research area aimed at making the decision-making processes of machine learning models transparent and interpretable. Explanations should satisfy two complementary criteria: they must be *faithful*, accurately reflecting the reasoning of the model, and *plausible*, making sense to humans who interpret them [371, 369]. Plausibility evaluation necessitates human domain knowledge. Balancing these two dimensions remains challenging, as methods that improve human comprehensibility do not not always guarantee a faithful reflection of the model's inner logic, and vice versa. A second, widely discussed challenge concerns the trade-off between *interpretability* and *predictive performance* [93]: complex, high-performing models are often opaque, whereas simpler, interpretable ones may be perceived as less expressive. Nonetheless, a growing body of research demonstrates that both tensions can be alleviated: models can achieve high accuracy while offering explanations that are both faithful to their computations and meaningful to human users.

While several taxonomies of XAI methods exist, here we adopt a widely accepted distinction into two macro-categories. The first includes *intrinsic* or *interpretable-by-design* approaches, in which models are inherently transparent.

Examples comprise architectures that learn disentangled latent representations [226], attention-based relevance mechanisms [361], and prototype- or concept-based structures that ground predictions in human-understandable semantic units [340, 166, 3, 44, 51]. The second category encompasses *post-hoc* approaches, which seek to explain model behaviour after training through methods such as attention visualization [81], game-theoretic attributions like SHAP [186], local surrogate models such as LIME [276], gradient-based saliency methods [302, 293, 305], and emerging neuro-symbolic strategies [8]. The approaches presented here represent only a subset of a broader and rapidly evolving landscape of XAI research.

Explainability research in text and vision has matured considerably, benefiting from the strong alignment between human perception and model representations, but it remains far less developed in audio and speech domains. Existing explainable AI systems for speech and audio remain predominantly *attribution-based*,[1] relying on saliency maps [302], attention weights [81], or numerical attribution scores [186, 319] to indicate which signal-level or spectral components influence model decisions. Yet such methods struggle with high-dimensional, non-linear architectures [93]. Linear proxy models like LIME [276] often yield inaccurate local approximations [281], and visualisation-based explanations [270, 302] fail to clarify how the model actually exploits the highlighted information. This limitation reflects a deeper mismatch: unlike words or pixels, acoustic representations lack direct perceptual or linguistic correlates, making interpretability in audio inherently difficult.

Recent progress in audio interpretability has shown that meaningful explanations can be achieved beyond traditional attribution-based techniques. Prototype- and concept-based approaches [386, 183, 240] have demonstrated that models can ground their reasoning in perceptually interpretable acoustic patterns, moving closer to how humans perceive and describe sound. Building on this direction, a complementary line of work has focused on *listenable explanations*, which make model reasoning perceptible through sound itself. SLIME [214] adapts LIME by segmenting spectrograms into time-frequency regions, while AudioLIME [108] extracts sources and assigns saliency scores. Listen-to-Interpret (L2I) [245] introduces a decoder that activates NMF components relevant to predicted classes; L-MAC [242] simplifies this with a faithfulnes-oriented loss and no pre-trained dic-

[1] *Attribution-based* methods estimate the contribution or importance of each input element (e.g., time frame, frequency bin, or token) to the model's prediction.

tionaries, and LMAC-ZS [241] extends it to zero-shot audio classification. However, L-MAC's reliance on magnitude-STFT representations leads to reconstruction artefacts. Recent advances in deep learning-based source separation, achieving Signal-to-Noise Ratios above 20 dB on benchmarks like WSJ0-2Mix [112] and Libri2Mix [60], now enable clearer and more natural listenable explanations, paving the way for perceptually faithful and human-centred interpretability in audio.

These challenges become even more pronounced in multimodal architectures, where audio must interact with textual or visual representations that differ both structurally and perceptually. Tools such as M2Lens [353] and MultiViz [172], which support the analysis of text, audio, and visual modalities, provide visualization-driven insights into cross-modal interactions. However, when applied to speech, these systems rely on spectrogram-based visualizations that are difficult to interpret and do not align with how humans perceive or describe sound. While prototype-, concept-, and listenable-explanation methods show promise in audio-only scenarios, they remain underexplored in multimodal contexts. Addressing this requires frameworks that align model representations with perceptually grounded auditory concepts rather than abstract features, enabling explanations that remain meaningful across modalities.

6.3 LMAC-TD: Producing Time Domain Explanations for Audio Classifiers

This section presents our work on generating listenable explanations for audio classifiers, conducted in collaboration with Francesco Paissan, Mirco Ravanelli, and Cem Subakan [194].[2]

Key Contribution: We introduce LMAC-TD (Listenable Maps for Audio Classifiers - Time Domain), a post-hoc explanation method that produces interpretations directly in the time domain, enabling domain experts to hear exactly which acoustic properties drive classifier decisions.[a] Building upon L-MAC [242], which generates explanations in the magnitude-STFT domain, LMAC-TD incorporates SepFormer [317], a transformer-based source separation architecture, to eliminate phase estimation artefacts and generate higher-quality listenable explanations while maintaining faithfulness to classifier decisions. The work addresses a critical gap in explainable AI for audio by establishing that listenability—the property of explanations being evaluable through direct auditory perception—is both measurable and achievable without sacrificing faithfulness. Through systematic evaluation on environmental sound classification using CNN14 classifiers [149] trained on the ESC50 dataset [251], we demonstrate that LMAC-TD achieves comparable or superior faithfulness metrics to existing explanation methods while significantly improving perceived audio quality as validated through user studies with 19 participants.

[a]Code available at: `https://github.com/fpaissan/lmac-td-code`

Evaluation across in-domain settings and multiple out-of-domain conditions, including audio mixtures, white noise contamination, and speech contamination, confirms robustness. We further show through ablation studies that the convex combination parameter α, which balances contributions from classifier representations and encoded audio, provides a tunable mechanism for navigating the faithfulness-audio quality trade-off. These results establish that perceptually-aligned explanations can serve as practical instruments for domain expert validation, providing a proof of concept that extends to speech-based multimodal

[2]© 2025 IEEE. Reprinted, with permission, from [194].

systems where combining listenable audio explanations with readable text explanations enables holistic interpretability.

6.3.1 Methodology

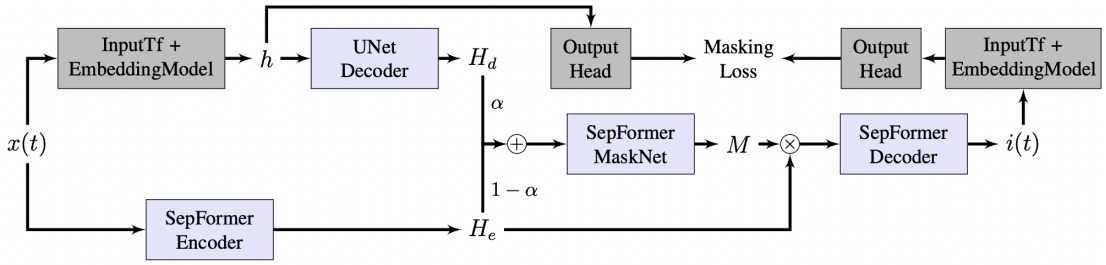


Figure 6.1: *The pipeline of LMAC-TD: The time domain input signal $x(t)$ is fed into embedding model to obtain the classifier representations h. These classifier representations are then fed into the UNet Decoder, which consists of a series of conv-transpose operations and skip connections. The UNet Decoder produces a latent representation that is of the same shape as the output of the SepFormer Encoder. This representation H_d is then combined with the encoder output H_e and fed into the SepFormer MaskNet. The output of MaskNet M is then element-wise multiplied and fed into the SepFormer Decoder to produce the time domain interpretation $i(t)$. During training, this time domain interpretation is then fed back into the classifier to calculate the masking loss. Light-blue boxes represent modules whose parameters are updated. Figure taken from [194]. © 2025 IEEE.*

We describe the pipeline of LMAC-TD in Figure 6.1. The goal of this model is to produce a time-domain interpretation signal $i(t)$ for a time-domain input signal $x(t)$ that goes through the classifier $f(\cdot)$, composed of an embedding model Emb$(\cdot)$ and a classification head (OutHead in the diagram). We enhance the L-MAC architecture by adding the SepFormer's MaskNet, Encoder, and Decoder on top of the UNet Decoder.

First of all, the classifier representations are obtained by passing the input signal $x \in \mathbb{R}^T$ first through an input transformation InputTf$(\cdot)$ (e.g. STFT, Mel Spectrogram extraction), and then through the embedding model of the pre-trained classifier, Emb$(\cdot)$. We obtain a set of classifier representations such that,

$$h = \text{Emb}(\text{InputTf}(x)). \tag{6.1}$$

For our experiments, InputTf$(\cdot)$ represents the Mel-spectrogram computation, while Emb$(\cdot)$ refers to a CNN14 encoder [149]. For h, we use the last 4 representations of the CNN14 encoder. The representations h are then fed into a UNet

decoder $U(\cdot)$, which is functionally similar to the decoder architecture in the L-MAC paper [242] (Appendix A), but with adjusted strides and kernel sizes. We obtain the representation $H_d \in \mathbb{R}^{K \times T'}$ such that,

$$H_d = U(h). \tag{6.2}$$

In order to take into account contributions from both classifier representations and the encoded input audio directly, we combine H_d with the output of the SepFormer encoder [318] $H_e \in \mathbb{R}^{K \times T'}$. The convex combination of H_e and H_d is then fed into the SepFormer MaskNet to obtain the mask $M \in \mathbb{R}^{K \times T'}$ such that,

$$M = \text{MaskNet}\Big(\alpha H_d + (1-\alpha)H_e\Big), \tag{6.3}$$

with $\alpha \in [0, 1]$ a scalar weighting parameter. We observe that calculating this combination through alpha allows the model to reach better faithfulness-audio quality trade-off. Finally, the mask M is element-wise multiplied with the Sepformer Encoder output H_e, and fed into SepFormer Decoder to obtain the time domain interpretation $i \in \mathbb{R}^T$ such that,

$$i = \text{SepFormerDecoder}\Big(M \odot H_e\Big). \tag{6.4}$$

The loss function is then calculated by comparing the classification results for the input signal x, the interpretation i, and the masked-out version of the interpretation signal i_{out}, which is defined as,

$$i_{\text{out}} = \text{SepFormerDecoder}\Big((1-M) \odot H_e\Big). \tag{6.5}$$

All in all, the training loss function is calculated following the L-MAC approach,

$$\begin{aligned} \min_{\theta} \lambda_{\text{in}} d\Big(f\Big(\text{InputTf}(x)\Big) \parallel f\Big(\text{InputTf}(i)\Big)\Big) \\ - \lambda_{\text{out}} d\Big(f\Big(\text{InputTf}(x)\Big) \parallel f\Big(\text{InputTf}(i_{\text{out}})\Big)\Big) + \lambda_{\text{reg}} R(i), \end{aligned} \tag{6.6}$$

where $d(\cdot||\cdot)$ is a divergence measure which measures the discrepancy between the classifier output $f(\text{InputTf}(x))$, and the classification of the interpretation

Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
L-MAC	36.25	**1.15**	23.50	0.20	0.42	0.47	**4.71**
L-MAC, FT, $\lambda_g = 4$	32.37	1.98	18.74	0.21	0.41	0.43	5.20
LMAC-TD, $\alpha = 1.00$ (ours)	66.00	2.62	22.39	**0.42**	0.87	**0.86**	10.50
LMAC-TD, $\alpha = 0.75$ (ours)	**69.75**	2.10	**28.07**	**0.42**	**0.91**	**0.86**	10.53
LMAC-TD, $\alpha = 0.00$ (ours)	46.50	5.55	11.86	**0.42**	0.86	0.80	10.88

Table 6.1: *In-domain quantitative evaluation on the ESC50 dataset. Our results reveal that LMAC-TD achieves higher faithfulness scores (AI, AG, FF, Fid-In, SPS) compared to other methods.*

$f(\text{InputTf}(i))$. In our experiments, we use Cross-Entropy as $d(\cdot||\cdot)$. The second term tries to make sure that the masked-out interpretation signal i_{out} minimises its similarity to the input signal x in terms of the classifier output. The coefficients λ_{in} and λ_{out} are used to control the relative strengths of the aforementioned mask-in and mask-out loss terms. Finally, the last term $R(i)$ is a regularisation term that avoids trivial solutions.[3] In our experiments, we applied an ℓ^1 penalty to the magnitude-STFT representation derived from the interpretation signal such that,

$$R(i) = \|\text{STFT}(i)\|_1. \tag{6.7}$$

6.3.2 Experiments

In Sec. 6.3.2 we detail the experimental setup used to validate LMAC-TD. We performed a quantitative analysis to establish the faithfulness of LMAC-TD-generated explanations, which we describe in Sec. 6.3.2. Additionally, we conducted a user study to evaluate the user preference and present the results in Sec. 6.3.2.

Setup

Metrics. We use several quantitative metrics to compare LMAC-TD with state-of-the-art approaches. Specifically, we adopt the evaluation metrics employed in

[3] In case there are no constraints, the learning objective in Eq. 6, is minimised by an all-ones mask. Of course, an all-ones mask is not informative in terms of providing insightful explanations, and therefore should be avoided.

the original L-MAC paper [242]. We use Average Increase (AI), which measures the increase in the classifier's confidence for the interpretation with respect to the input sample, Average Drop (AD), which measures the confidence drop when removing the interpretation from the input (i.e. feeding i_{out}), and Average Gain (AG), similar to AI. Beyond these metrics, we use the Faithfulness (FF) metric defined in the L2I paper [245] and the input fidelity (Fid-In) metric defined in the PIQ paper [243]. We also use the Sparseness (SPS) [38] and Complexity (COMP) [26] metrics to evaluate the conciseness of the explanations. Due to space constraints, we invite the reader to refer to the respective papers for an in-depth presentation of the metrics.

Implementation. For all experiments, we use a CNN14 classifier [149] pre-trained on the VGGSound dataset [45]. This classifier is then fine-tuned on the ESC50 dataset [251] with WHAM! noise [360] augmentation to mimic real-world conditions. It is trained on folds 1, 2, and 3, achieving 75% and 78% accuracy on folds 5 and 4, respectively. LMAC-TD is later trained on ESC50 with WHAM! noise augmentation. We refer to this data configuration when evaluating the explanations in the In-Domain setting. The UNet decoder employs a series of transposed convolution layers to align the last four latent representations with the encoder output H_e shape.

We benchmarked LMAC-TD with different α values to assess their impact. The α parameter primarily analyzes the tradeoff between classifier representations and SepFormer encoder output. The remaining hyperparameters were set as $\lambda_{\text{in}} = 5$, $\lambda_{\text{out}} = 0.2$, and $\lambda_{\text{reg}} = 6$, chosen through manual search since the model is not highly sensitive to these values. This combination was selected based on qualitative evaluation of audio quality, taking into account the faithfulness of the explanations.

Our implementation is based on the L-MAC implementation present in SpeechBrain 1.0 [271] and is available through our companion website[4].

Evaluation Data. We conduct experiments to evaluate explanation faithfulness on In-Domain (ID) and Out-of-Domain (OOD) data. For the OOD setting, following L-MAC's procedure, we create mixtures of ESC-50 data, which contain samples for environmental sound classification, with the dominant class at 5 dB SNR. We generate mixtures from folds 4 and 5. We also test LMAC-TD on audio contaminated with white noise and speech, using 3dB SNR mixtures and samples

[4] https://francescopaissan.it/lmac-td

Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Saliency	0.62	31.73	0.07	0.06	0.12	0.76	11.06
Smoothgrad	0.12	31.84	0.00	0.06	0.13	0.83	10.66
IG	0.37	31.15	0.03	0.12	0.26	0.87	10.22
L2I	5.00	25.65	1.00	0.20	0.35	0.52	10.99
GradCAM	14.12	17.62	7.46	0.25	0.00	0.91	9.66
Guided GradCAM	0.00	31.74	0.00	0.07	0.11	0.89	10.24
Guided Backprop	0.63	31.73	0.07	0.06	0.11	0.76	11.06
SHAP	0.00	31.81	0.00	0.07	0.14	0.84	10.58
L-MAC	**60.63**	4.82	**35.85**	0.39	0.81	**0.94**	**9.61**
L-MAC FT, $\lambda_g = 4$	50.75	6.73	26.00	0.39	0.78	0.84	10.51
LMAC-TD, $\alpha = 1.00$ (ours)	56.75	3.62	16.84	**0.42**	**0.88**	0.89	10.36
LMAC-TD, $\alpha = 0.75$ (ours)	59.50	**3.42**	21.22	0.41	0.87	0.88	10.35
LMAC-TD, $\alpha = 0.00$ (ours)	39.88	7.60	9.30	0.42	0.82	0.83	10.69

Table 6.2: *Out-of-Domain Quantitative Evaluation for the ESC50 Dataset. In out-of-distribution conditions, LMAC-TD achieves results comparable to L-MAC.*

from the LJSpeech dataset [120].

Faithfulness Evaluation

A very critical aspect of evaluating LMAC-TD explanations is establishing that they faithfully follow the classifier, meaning they effectively highlight portions of the input relevant for the classifier. We compare LMAC-TD with several other saliency-based methods in the literature. Namely, we compared with gradient-based methods such as saliency maps [302], GradCAM [293], SmoothGrad [305], IntegratedGradients [319], GuidedBackProp [312], the decoder-based audio-specific explanation method Listen-to-Interpret (L2I) [245], and we also include SHAP [186].

In Table 6.1, we report the quantitative results obtained for the ID experiments with three different LMAC-TD configurations. We observe that, with $\alpha = 0.75$, LMAC-TD outperforms standard L-MAC in all faithfulness metrics except AD, which remains comparable. We note that we have calculated the COMP metric (a complexity metric) on the magnitude-STFT representations of the explanation signal i, which is produced in the time domain. This tends to give large COMP metrics for LMAC-TD, implying that LMAC-TD explanations are complex (high entropy - uniformly distributed in the STFT domain). However, LMAC-TD explanations remain understandable, and the explanations are still preferable for the users over the alternatives as we show in Section 6.3.2 with a user study.

We note that the results show that lowering α reduces the faithfulness met-

Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
	Classification on ESC50, White Noise Contamination, 38.6% accuracy						
L2I @ 0.2	0.00	19.41	0.21	0.11	0.04	36.62	**7.32**
L-MAC	**83.62**	**1.50**	**56.12**	0.33	**0.86**	**0.92**	10.03
LMAC-TD, $\alpha = 1$ (ours)	55.00	4.32	16.15	0.35	0.81	0.82	10.75
LMAC-TD, $\alpha = 0.75$ (ours)	63.50	3.06	24.57	0.35	0.83	0.84	10.71
LMAC-TD, $\alpha = 0.00$ (ours)	33.50	10.03	6.83	**0.36**	0.71	0.75	11.09
	Classification on ESC50, LJSpeech Contamination, 79.3% accuracy						
L2I @ 0.2	1.75	29.49	0.27	0.15	0.18	0.79	**9.56**
L-MAC	**70.75**	**2.73**	**39.64**	0.33	0.83	**0.93**	9.70
LMAC-TD, $\alpha = 1.00$ (ours)	59.13	3.05	17.18	0.43	**0.90**	0.88	10.33
LMAC-TD, $\alpha = 0.75$ (ours)	62.63	2.95	21.57	0.43	**0.90**	0.88	10.33
LMAC-TD, $\alpha = 0.00$ (ours)	44.25	6.50	9.88	**0.44**	0.87	0.83	10.69

Table 6.3: *Additional results obtained on the ESC50 dataset with White Noise and LJSpeech contamination.*

rics. This makes sense since a larger α more strongly incorporates the classifier representations, increasing the classifier influence on the explanations.

In Table 6.2, we compare the results obtained on audio mixtures created with ESC50 data. Unlike the ID case, LMAC-TD with $\alpha = 0.75$ outperforms both L-MAC and STFT-masking approaches in AD, FF, and Fid-In. With regards to the effect of α, we observe that in general the faithful metrics are better as expected with larger α, however in terms of FF, $\alpha = 0$ remains comparable. These results suggest that the explanations generated with LMAC-TD are well-aligned with the classifier in this OOD setting as well.

Finally, Table 6.3 presents the results on the ESC50 dataset with white noise and LJSpeech contamination[5] to explore more OOD settings. In these settings, LMAC-TD achieves the best overall faithfulness scores with $\alpha = 0.75$. Compared to standard L-MAC, we observe a drop in terms of the performance in some metrics, such as AI, AD AG, however in terms of FF and Fid-In LMAC-TD achieves better scores. We also would like to note that, similar to the ID setting, in all OOD settings, larger α typically results in better faithfulness scores.

Subjective Evaluation

To determine the perceived quality of the generated interpretations, we conducted a user study with 19 participants using the WebMushra interface [290], following the evaluation pipeline used for L-MAC [242]. We presented nine audio samples

[5]In the table, we have included only the top two methods for comparison; the complete set of results is available on the companion website.

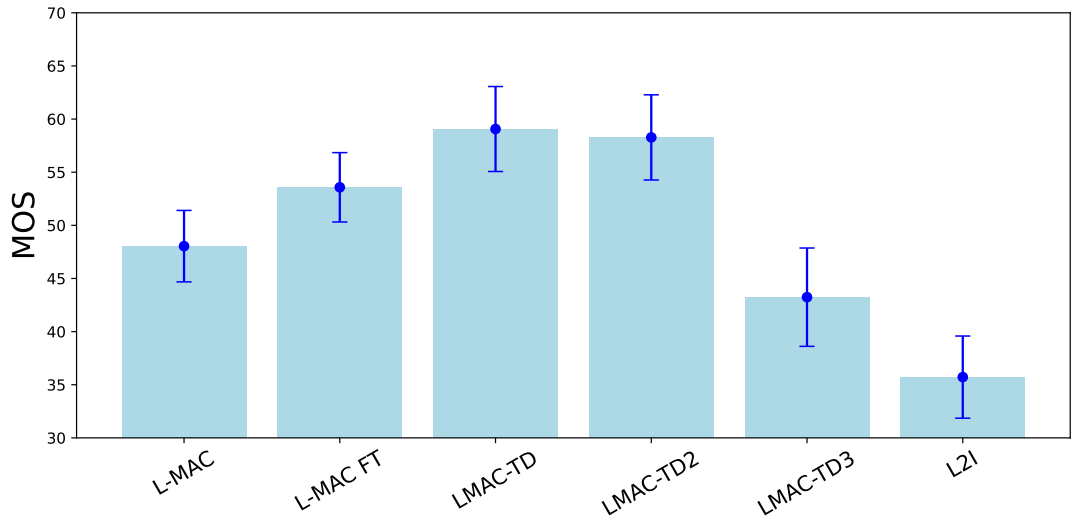


Figure 6.2: *Mean-Opinion Score (MOS) for the interpretations. Confidence intervals at* 0.95 *are reported as error bars (in blue). In the Figure, LMAC-TD corresponds to the configuration with* $\alpha = 1$*, LMAC-TD2 to the configuration with* $\alpha = 0.75$*, and LMAC-TD3 to the configuration with* $\alpha = 0$*. Figure taken from [194]. © 2025 IEEE.*

to the users, and we asked the users to assign a score (1 to 100) based on (i) how well the interpretation corresponds to the part of the input audio associated with the predicted class and (ii) the perceived audio quality. These criteria are not rated separately but are combined into a single metric reflecting the overall user perception. We present the outcome of the user study in Fig. 6.2 with the confidence intervals at 0.95. LMAC-TD outperforms the other methods when $\alpha \in \{1, 0.75\}$. With $\alpha = 0$, LMAC-TD obtains a Mean-Opinion Score (MOS) comparable to the original L-MAC, while still outperforming L2I.

We also observe that LMAC-TD is preferred over LMAC-FT (the fine tuned version of LMAC [242]). We note that LMAC-TD simplifies the pipeline by removing the need for interpreter finetuning, while improving perceived audio quality over LMAC-FT. The audio samples from the user study are accessible through our companion website[1] together with the per-sample MOS.

In summary, faithfulness and qualitative evaluations demonstrate that LMAC-TD with $\alpha = 1$ and $\alpha = 0.75$ both offer better subjective audio quality compared to the baselines, while not compromising significantly on faithfulness.

6.3.3 Findings

We presented a novel approach, called LMAC-TD, to generate listenable explanations directly in the time domain. LMAC-TD builds on top of recent literature in interpretability and source separation to enhance the audio quality of the explanation. We proposed to include a time-domain source-separation architecture to decode listenable explanations from the classifier representations. Through an extensive experimental analysis, we conclude that LMAC-TD outperforms the other methods in terms of perceived audio quality and achieves better or comparable results in faithfulness metrics.

6.4 Conclusion

This chapter addresses the challenge of explainability in speech-based multimodal systems by introducing *listenability* as a critical property for audio explanations and demonstrating its feasibility through LMAC-TD, a time-domain method for generating perceptually aligned interpretations. It establishes that explanation properties must be grounded in the perceptual characteristics of each modality and how humans naturally process them. Given the asymmetry between our ability to describe visual and linguistic information and our limited vocabulary for auditory experiences, we argue that effective audio explanations must operate at the level of human auditory perception rather than technical feature representations.

Through systematic evaluation on environmental sound classification, LMAC-TD achieves comparable or superior faithfulness to existing methods while significantly improving perceived audio quality. The time-domain formulation, incorporating SepFormer components, eliminates phase estimation artifacts present in frequency-domain approaches, producing clearer and more interpretable explanations. User studies with 19 participants confirm a consistent preference for LMAC-TD, validating that improved listenability enhances practical usability. Ablation studies further show that the convex combination parameter α controls the faithfulness-audio quality trade-off, with $\alpha = 0.75$ achieving optimal balance across varied conditions. These results establish listenability as both measurable and attainable, transforming audio explainability from technical attribution into a perceptually grounded interpretability framework.

The implications of this work extend across the multimodal domains explored in this thesis. In argument mining, listenable explanations isolating paralinguistic cues such as stress, pauses, emotional tone, combined with readable text explanations highlighting claims and premises could help scholars validate computational predictions against rhetorical theory. In clinical decision support, explanations that present both acoustic and textual evidence, such as flattened prosody, reduced speech rate, or specific linguistic markers, would allow clinicians to evaluate model reasoning against their diagnostic expertise. In music information retrieval, distinguishing audio-based from lyrics-based similarity through separate listenable and readable explanations would support copyright analysis and creative workflows by clarifying modality contributions.

Despite these advances, several limitations guide future work. LMAC-TD is

currently validated only on environmental sounds; extending it to speech tasks will test whether listenability principles transfer to human vocalization, where linguistic and paralinguistic features interact. Integrating listenable and readable explanations within synchronized multimodal interfaces represents the next step toward holistic interpretability. Future evaluation should also move beyond user preference to task-based studies measuring whether experts perform better with perceptually aligned explanations. Finally, developing intrinsically interpretable architectures that produce listenable prototypes during training could unify transparency and performance.

Beyond the audio domain, the broader challenge is to establish *perceptual alignment* as a general design principle for explainable multimodal AI. Identifying modality-specific explanation properties, such as tactile realism for haptics or temporal coherence for video, would enable explanation frameworks that align with human perceptual and cognitive capacities. Standardized evaluation protocols that measure perceptual alignment through cognitive load or task performance could then guide the selection of explanation types across applications.

In conclusion, this chapter demonstrates that perceptually aligned audio explanations are achievable without compromising faithfulness, establishing *listenability* as a measurable and meaningful property for XAI. By focusing on human auditory perception rather than abstract features, LMAC-TD transforms explainability from a diagnostic tool into a practical instrument for domain expert validation and human-AI collaboration. The convergence of needs across argument mining, clinical decision support, and music analysis underscores the importance of perceptually grounded explainability as a central component of future multimodal AI systems.

CHAPTER 7

Conclusion

In an era when we increasingly expect AI systems to act as collaborators and companions, we must equip them with the ability of communicating and understanding communication not only in terms of content, but also expression in a broader sense. Yet a discontinuity persists between human expression and machine communication. Bridging this gap requires multimodal AI capable of mirroring the richness, ambiguity, and emotional depth of human interaction, grounding computational intelligence in the same perceptual and expressive dimensions that define human communication.

This thesis advances the understanding of human expression through multimodal artificial intelligence (AI), investigating how speech, text, and other complementary signals can be integrated, interpreted, and explained in human-centred contexts. By systematically exploring three distinct yet interconnected domains, namely political discourse analysis, clinical decision support, and music information retrieval, this work establishes the methodological foundations and ethical standards necessary for trustworthy multimodal learning across sensitive and creative environments.

The research defines a coherent trajectory spanning the entire multimodal AI pipeline—from data creation to interpretability. Chapter 3 demonstrates that argument mining benefits from the integration of prosodic and paralinguistic cues with textual content. The introduction of MM-USED and MM-USED-fallacy establishes the first large-scale multimodal resources for argument mining. Evaluation conducted on these corpora, together with prior datasets, confirms that multimodal models achieve superior performance over text-only approaches in detecting argumentative structures and fallacies. The accompanying toolkit, MAMKit, and the organization of the MM-ArgFallacy2025 shared task provide an open, reproducible infrastructure for community-wide experimentation.

In Chapter 4, the focus shifts to the clinical domain, where multimodal AI must meet stringent standards of accuracy, interpretability, and ethical responsibility. We begin with a detailed case study that informs the development of a comprehensive framework for responsible data collection and curation. Applying this framework to a systematic review of clinical speech datasets uncovers substantial ethical shortcomings in consent, transparency, and demographic representation. After exploring the data-related challenges through this framework, we turn to the methodological problem of multimodal integration. The proposed TRI-DEP architecture demonstrates the diagnostic potential of fusing speech, text, and EEG signals for depression detection, while ensuring rigorous subject-level validation. Finally, we benchmark post-hoc explainability methods on spectrogram-based models for Parkinson's disease detection from speech, showing that while gradient- and perturbation-based approaches achieve high technical faithfulness, they fail to produce explanations that align with clinical reasoning. This highlights the broader need for interpretability methods that genuinely support expert understanding and trust. Together, these contributions show that accurate and responsible clinical AI emerges from the integration of ethical practices, robust multimodal modelling, and transparent reasoning, while also highlighting current limitations and pointing to future directions for more interpretable and clinically meaningful multimodal systems.

Continuing the exploration of multimodal learning across domains, Chapter 5 examines the creative context of MIR, addressing the challenges of data scarcity and multimodal integration. It introduces WEALY, the first fully reproducible end-to-end pipeline that learns lyrics-aware audio representations directly from raw music without requiring pre-existing transcriptions. By adapting Whis-

per's decoder embeddings through contrastive learning, WEALY overcomes the severe data scarcity that has long limited multimodal MIR. Systematic evaluation across three diverse datasets shows that the model consistently outperforms transcription-based approaches, and simple late fusion with audio-content baselines yields substantial gains in version identification accuracy, confirming that lyrics and audio encode complementary and orthogonal information. Beyond performance, WEALY provides a degree of interpretability with separate similarity scores for lyrics and audio, enabling analysis of modality contributions in contexts such as copyright verification and stylistic comparison. By combining interpretability with performance, this work illustrates how multimodal integration can advance both analytical and creative objectives, supporting copyright analysis, stylistic exploration, and artistic understanding, while highlighting opportunities for more sophisticated fusion and explainability methods in future research.

Finally, Chapter 6 introduces the concept of perceptually aligned explainability, reframing interpretability as a property rooted in human perception rather than technical abstraction. Through the LMAC-TD model, this work defines listenability as a measurable dimension of audio explanations, showing that models can produce faithful and perceptually meaningful interpretations. This concept establishes the foundation for explainability methods that engage users through the same perceptual channels by which humans naturally interpret information.

Together, these contributions outline a comprehensive direction for multimodal AI, demonstrating how advances in data, fusion, and interpretability can jointly enhance performance and transparency. The resulting approach generalises across domains, providing a foundation for future AI systems that are accurate, transparent, and aligned with human expression.

7.1 Summary of Contributions

In this section, we summarise how the research questions presented in Chapter 1 are addressed throughout the thesis.

RQ1 — Multimodal Necessity

This thesis demonstrates empirically that multimodal approaches consistently surpass unimodal systems across all investigated domains. Speech, text, music audio, lyrics, and other complementary modalities such as EEG encode distinct yet syn-

ergistic cues — semantic, prosodic, acoustic, and neurophysiological — which, when combined, provide a more faithful representation of human communication and expression (see Chapters 3, 4, 5).

RQ2 — Cross-Domain Challenges

Through comparative analysis across political, clinical, and creative domains, this work identifies three central challenges: data collection and alignment, multimodal fusion, and interpretability. It introduces domain-independent methodologies for addressing each of them, including ethical dataset design and curation, and perceptually grounded explainability (see Chapters 3-6).

RQ3 — Systematic Solutions

Systematic solutions to multimodal learning are realized through the release of open and reproducible frameworks, including, but not limited to, MAMKit, TRI-DEP, WEALY, and LMAC-TD. Together, these tools enable benchmarking, transparency, and reproducibility across multimodal AI research (see Chapters 3-6).

RQ4-RQ5 — Multimodal Argument Mining

By introducing MM-USED and MM-USED-fallacy, the thesis aligns textual argument mining resources with speech data and demonstrates that multimodal models outperform text-only approaches in argument structure detection and fallacy classification. These contributions uncover the role of prosody and paralinguistic cues in persuasive discourse (see Chapter 3).

RQ6-RQ8 — Multimodal Clinical Decision Support

This research proposes an ethical framework for clinical speech data, develops the TRI-DEP architecture for fusing speech, text, and EEG signals in depression detection, and exposes the limitations of current explainability techniques in clinical interpretability. Together, these contributions establish guidelines for developing trustworthy and transparent multimodal clinical AI (see Chapter 4).

RQ9-RQ10 — Multimodal Music Information Retrieval

Through the introduction of WEALY, the first end-to-end pipeline for lyrics-aware audio representation, this work addresses the challenge of large-scale lyrics

data scarcity and demonstrates that late fusion between audio and lyrics yields measurable improvements in version identification. These advances contribute to both performance and interpretability in multimodal MIR (see Chapter 5).

7.2 Discussion and Future Work

By establishing a foundation for robust, effective, and explainable multimodal AI, this thesis opens several promising research directions.

The findings presented here highlight that multimodality is not merely advantageous but essential: many aspects of meaning, emotion, and intent remain inaccessible when relying on a single modality. This recognition points toward extending the proposed methodologies to additional sensory channels, particularly vision, to enable richer and more comprehensive multimodal reasoning. In political discourse (see Chapter 3), integrating gesture, gaze, and facial expression with speech and text could deepen our understanding of persuasion, deception, and emotional framing. In clinical contexts (see Chapter 4), fusing speech with facial expressions and neurophysiological signals may yield more comprehensive biomarkers and improve interpretability in real-world decision support. In creative domains (see Chapter 5), combining audio and lyrics with visual representations of performance or musical notation could support stylistic exploration and copyright validation.

Another key finding of this thesis is the identification of *listenability* as a cornerstone of explainability for speech-based systems. The results presented in Chapter 6 show that perceptually aligned explanations can be both faithful and accessible. Future work could expand this concept to speech-related applications where auditory cues are essential—for instance, affect recognition, conversational understanding, or early clinical screening—and establish standardised evaluation frameworks that assess both faithfulness and perceptual quality.

Recent advances show that the next frontier in AI is the emergence of multimodal LLMs, capable of processing and generating information across text, vision, and speech [372]. Models such as GPT-5.2 [235] and Gemini 3 [96] exemplify this evolution, extending linguistic reasoning into perceptual understanding. Yet, as these systems grow in scale and capability, critical challenges persist: the responsible creation and use of multimodal datasets in sensitive domains, and the lack of interpretability that limits their trustworthy deployment. The findings of this thesis address precisely these open points, demonstrating that it is possible to

advance interpretability and ethical multimodal integration in ways that make such technologies viable for contexts like political discourse analysis, clinical assessment, and creative expression.

Recent work also suggests that argument mining is becoming a powerful framework for enhancing reasoning in LLMs [248, 174, 168]. Fallacy detection and argumentative structure analysis offer promising means to evaluate and guide model reasoning. Our multimodal approach to fallacy identification and argumentative understanding provides a pathway for extending these ideas to LLMs, enabling systems that reason contextually and multimodally.

In this sense, the progress achieved in this thesis contributes to the broader transformation of multimodal AI, bringing us closer to intelligent systems that can see, listen, and understand as we do, and that can engage with humans not as opaque engines of computation but as transparent partners in reasoning and expression.

Acknowledgements

I wish to express my deepest gratitude to my advisor, Professor Paolo Torroni, for his guidance, trust, and continuous support throughout these years, and for opening my eyes to new worlds and perspectives. I am also grateful to Professors Federico Chesani and Simone Martini for their valuable feedback and mentorship. I warmly thank Professors Cem Subakan and Mirco Ravanelli at MILA - Québec AI Institute, and Dr. Joan Serrà at Sony AI, for their supervision during my visiting research periods. I am grateful to Professors Subakan and Ravanelli for sharing their expertise in speech processing and explainable AI, and to Joan Serrà for introducing me to music information retrieval. Finally, I sincerely thank my family for their unwavering support. I am also grateful to my father for his insights on the clinical aspects of this work.

Parts of this research have been supported by the European Commission's NextGenerationEU programme through PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", and by the EU H2020 ICT48 project "Humane AI Net" under contract #952026.

The depression detection research presented in Chapter 4 uses data from the MODMA dataset [36], provided by the Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, China. We gratefully acknowledge the data contributors for granting access. This study was conducted retrospectively using human subject data made available under the dataset's terms of use. Access is granted by the data owners upon request; we do not redistribute any data.

According to the terms specified by the dataset providers, separate ethical approval was not required for our analyses. All experiments were carried out in full compliance with these conditions.

Part of this research was carried out using computational resources provided by ISCRA on the LEONARDO supercomputer, owned by EuroHPC Joint Undertaking and hosted by CINECA, Italy.

List of Figures

3.1 The proposed schema for MAM. 53
3.2 Distribution of the character-level coverage ratio (fallacy span length / reference sentence length) per fallacy category, computed on the 1,891 samples in MM-USED-fallacy. The green zone indicates samples where the span closely matches the sentence boundaries; the red zone highlights cases where a short span was mapped to a much longer sentence, potentially diluting the audio signal. Each violin is annotated with η, the median of the coverage ratio for that category, i.e. the value below which half of the samples fall: a low η indicates that, for a typical sample in that category, the fallacy span covers only a small portion of the reference sentence. 67
3.3 Overall architecture of MAMKit. 73
3.4 The CSA model architecture. 79
3.5 The Ensemble model architecture. 80
3.6 The Mul-TA model architecture. 81

4.1 Experimental Framework for Multimodal Depression Detection. . . 148

4.2 Analysis of the correlation between explanations overlap among interpretability techniques and faithfulness metrics. Combining interpretability techniques is most effective when there is already significant overlap between attribution masks. Figure taken from [195]. © 2025 IEEE. 158

4.3 Explanations generated for a PD sample correctly classified by HuBERT. The explanations highlight different portions of the spectrogram, suggesting that understandability is difficult to achieve in this setting. From left to right, top to bottom: original sample, Integrated Gradients, Guided Backprop, and Guided GradCAM. Figure taken from [195]. © 2025 IEEE. 159

5.1 Overview of the WEALY architecture. Stage 1 extracts lyrical latents with Whisper, while stage 2 learns contextualized representations with a transformer encoder and projects them into a compact embedding space. 177

6.1 The pipeline of LMAC-TD: The time domain input signal $x(t)$ is fed into embedding model to obtain the classifier representations h. These classifier representations are then fed into the UNet Decoder, which consists of a series of conv-transpose operations and skip connections. The UNet Decoder produces a latent representation that is of the same shape as the output of the SepFormer Encoder. This representation H_d is then combined with the encoder output H_e and fed into the SepFormer MaskNet. The output of MaskNet M is then element-wise multiplied and fed into the SepFormer Decoder to produce the time domain interpretation $i(t)$. During training, this time domain interpretation is then fed back into the classifier to calculate the masking loss. Light-blue boxes represent modules whose parameters are updated. Figure taken from [194]. © 2025 IEEE. 197

6.2 Mean-Opinion Score (MOS) for the interpretations. Confidence intervals at 0.95 are reported as error bars (in blue). In the Figure, LMAC-TD corresponds to the configuration with $\alpha = 1$, LMAC-TD2 to the configuration with $\alpha = 0.75$, and LMAC-TD3 to the configuration with $\alpha = 0$. Figure taken from [194]. © 2025 IEEE. . 203

List of Tables

3.1 Corpora for MAM. For M-Arg, we also consider the corpus version in which samples have high annotation confidence, γ. 50
3.2 Average binary F1-score regarding the *claim* class on the test set of UKDebates. For each row, we report the best results in bold, second best results are underlined instead. 55
3.3 Average macro F1-score concerning the *attack* and *support* classes on the test set of M-Arg ($\gamma \geq 0.85$). For each row, we report the best results in bold, second best results are underlined instead. . . 56
3.4 Average macro F1-score on the test set of MM-USED. For each row, we report the best results in bold, second best results are underlined. 57
3.5 Ablation test regarding TA model configuration on the UKDebates test set. 58
3.6 Ablation test regarding TA model configuration on the M-ARG ($\gamma \geq 0.85$) test set. 58
3.7 Ablation for ASD task 59
3.8 Ablation for ACC task 59
3.9 Annotations in USED-fallacy. 63
3.10 Examples of annotated fallacies. 64
3.11 Number of samples in USED-fallacy and MM-USED-fallacy. . . . 65
3.12 Text and audio length distribution (unit: sentences). 65
3.13 Snippet sentence length distribution per category (unit: sentences). 66

3.14 Result on MM-USED-fallacy. AE: *Appeal to Emotion*, AA: *Appeal to Authority*, AH: *Ad Hominem*, FC: *False Cause*, SS: *Slippery Slope*, S: *Slogans*. In bold the best model, underlined the second-best model. 70
3.15 Multimodal models available in MAMKit. For each model, we summarize its text and audio encoding modules and its fusion strategy. *Concat*: Concatenation; *Avg*: Average; *Cross*: Crossmodal Attention. 78
3.16 Test classification performance on MAM datasets. For each multimodal model, we report its constituting text module (T_i) and audio module (A_j). 83
3.17 Available datasets for MM-ArgFallacy2025 shared task. 87
3.18 Secret test set statistics for Argumentative Fallacy Classification. . 88
3.19 Results for multimodal argumentative fallacy detection on political debates. For AFC, we report the macro F1-score, while for AFD, we report the binary F1-score. Best results per subtask are in **bold**, second best results are underlined. 89
3.20 Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking. 90

4.1 Checklist for Mental and Neurological Health Speech Datasets. . 127
4.2 Number of datasets that address each checklist item grouped by target issue. 128
4.3 Number of datasets that address each checklist item grouped by discourse genre. 133
4.4 Datasets addressing each checklist item grouped by data collection method. 136
4.5 Participant demographics in the MODMA dataset. *M*: Male, *F*: Female, *HC*: Healthy Control, and *MDD*: Major Depressive Disorder. 140
4.6 Shapes of speech feature matrices per recording r and at the subject level. 143

4.7 Results of baselines and unimodal models (macro-F1, mean ± std across 5 folds). The *Features* column indicates the input representation (handcrafted or learned); the *Detection Module* column indicates the classifier trained on top. For speech, the detection module comprises a segment-level encoder and a subject-level classifier, separated by →. In **bold**, the best performing configuration per modality. 146
4.8 Baseline and multimodal fusion results (F1-score, mean ± std across 5 folds). In **bold**, the best performing configuration per category. The overall best are additionally underlined. Fusion weights were optimised via grid search (step = 0.05). 147
4.9 Pairwise statistical significance tests for 12 selected model comparisons. 149
4.10 PD detection results averaged over 10-folds on s-PC-GITA. Mean value and standard deviation are reported. 155
4.11 Results of PD detection using saliency maps. 156
4.12 PD detection quantitative evaluation results of HuBERT Base model averaged over 10-folds on s-PC-GITA. 157

5.1 Main results: MAP on the three considered datasets. 183
5.2 Ablation study of WEALY on the SHS dataset. 184
5.3 Comparison of audio content-based MVI approaches with the proposed multimodal approach the SHS dataset. 184

6.1 In-domain quantitative evaluation on the ESC50 dataset. Our results reveal that LMAC-TD achieves higher faithfulness scores (AI, AG, FF, Fid-In, SPS) compared to other methods. 199
6.2 Out-of-Domain Quantitative Evaluation for the ESC50 Dataset. In out-of-distribution conditions, LMAC-TD achieves results comparable to L-MAC. 201
6.3 Additional results obtained on the ESC50 dataset with White Noise and LJSpeech contamination. 202

List of Abbreviations

A

ACC	Argumentative Component Classification. 49, 51, 82, 87, 96
AFC	Argumentative Fallacy Classification. 41, 42, 64, 68, 71, 82, 86–89, 92, 94, 220
AFD	Argumentative Fallacy Detection. 85–89, 92, 220
AI	Artificial Intelligence. 1, 2, 5–8, 10, 11, 13, 15, 16, 19, 37, 40, 44, 47, 99, 100, 103, 107, 108, 152, 161, 162, 167, 169, 206–212, I
AM	Argument Mining. 37–41, 43, 46, 49, 51, 54, 86, 87, 96
AO	Audio-Only. 49, 52, 55–57, 59
ARC	Argumentative Relation Classification. 49, 50, 82, 87, 96
ASD	Argumentative Sentence Detection. 49, 51, 82, 87, 96
ASR	Automatic Speech Recognition. 6, 28, 141, 150, 151, 167, 172

B

BiLSTM — Bidirectional Long Short-Term Memory. 52, 53, 68, 77, 79, 89–91, 286

C

CD — Claim Detection. 49, 50

CNN — Convolutional Neural Network. 52, 91, 108–110, 141, 144–147, 149, 279

D

DCT — Discrete Cosine Transform. 28

E

EEG — Electroencephalography. 1–3, 9, 10, 12, 16, 25, 32–36, 102, 104, 108, 109, 139–144, 146, 147, 149, 150, 161, 163, 208–210

G

GRU — Gated Recurrent Unit. 144–146

H

HC — Healthy Control. 140, 154, 155, 158, 220

I

IAA — Inter-Annotator Agreement. 50, 51, 88

L

LLM — Large Language Model. 173, 211

LSTM — Long Short-Term Memory. 108, 109, 141, 144–147, 149, 279

M

MAM — Multimodal Argument Mining. 39–41, 44–50, 52, 53, 59, 60, 72, 75, 78, 82–85, 93, 96, 97, 217, 219

MDD Major Depressive Disorder. 112, 124, 140
MFCC Mel-Frequency Cepstral Coefficients. 28, 29, 53, 77, 90, 91, 141, 143, 146
MHND mental health and neurological disorders. 99, 101, 105, 106, 108, 113, 114, 116, 117, 119–123, 138, 161
MIR Music Information Retrieval. 2, 6, 35, 36, 166, 168–171, 174, 176, 185, 186, 208, 209, 211
MLP Multilayer Perceptron. 53, 79, 144, 145, 179, 180, 183, 184
MM-USED MM-USElecDeb60To16. 13, 17, 41–43, 49–51, 54, 56, 57, 59, 60, 64, 82, 83, 86, 87, 96, 98, 208, 210, 219
MM-USED-fallacy MM-USElecDeb60To16-fallacy. 13, 17, 42, 43, 61, 64–67, 70, 71, 82, 83, 86–88, 94, 96, 98, 208, 210, 217, 219, 220
MMDL Multimodal Deep Learning. 34, 35, 39, 42, 46, 47, 59
MVI Musical Version Identification. 3, 5, 6, 166, 168–171, 173, 174, 176, 184, 185

N

NLP Natural Language Processing. 22, 37, 39, 47

P

PD Parkinson's Disease. 153–159, 218, 221
PTSD Post-Traumatic Stress Disorder. 112

R

RLHF Reinforcement Learning from Human Feedback. 26

S

SSL	Self-Supervised Learning. 153
SVM	Support Vector Machine. 45, 52, 55

T

TA	Text-Audio. 49, 52, 55–59, 219
TO	Text-Only. 49, 52, 55–57, 59, 278, 280, 281

U

USED	USElecDeb60To16. 40, 41, 51, 62
USED-fallacy	USElecDeb60To16-fallacy. 62–65, 68, 69, 71, 219

X

XAI	Explainable Artificial Intelligence. 17, 18, 28, 103, 110, 153, 157–159, 190, 193, 194, 206

Bibliography

[1] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. Method cards for prescriptive machine-learning transparency. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI, CAIN 2022*, pages 90–100. ACM, 2022.

[2] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 9(4):478–490, 2018.

[3] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 7786–7795, 2018.

[4] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and the Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, November 2020.

[5] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible

data curation. In *Advances in Neural Information Processing Systems*, volume 36, pages 55320–55360. Curran Associates, Inc., 2023.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[7] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *arXiv preprint*, arXiv:2312.11805, 2023.

[8] Prashani Jayasingha Arachchige, Bogdan Iancu, and Johan Lilius. A roadmap toward neurosymbolic approaches in ai design. *IEEE Access*, 13:174368–174392, 2025.

[9] R Oguz Araz, Xavier Serra, and Dmitry Bogdanov. Discogs-VI: A musical version identification dataset based on public editorial metadata. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[10] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John T. Richards, Jason Tsay, and Kush R. Varshney. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.*, 63(4/5):6:1–6:13, 2019.

[11] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representa-

tions. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[12] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *arXiv preprint*, arXiv:2212.08073, 2022.

[13] Andrew Bailey and Mark D. Plumbley. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600, 2021.

[14] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493, 2023.

[15] Marieke Bak, Vince Istvan Madai, Marie-Christine Fritzsche, Michaela Th Mayrhofer, and Stuart McLennan. You can't have AI both ways: balancing health data privacy and access fairly. *Frontiers in genetics*, 13:929453, 2022.

[16] Maximilian Balluff, Maximilian Auch, Peter Mandl, and Christian Wolff. Lyriccovers 2.0: An enhanced dataset for cover song analysis. *IADIS International Journal on WWW/Internet*, 22(2):75–92, 2024.

[17] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

[18] James T Becker. The natural history of alzheimer's disease. *Arch. Neurol.*, 51(6):585, June 1994.

[19] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *arXiv preprint*, arXiv:1807.03418, 2023.

[20] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[21] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[22] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[23] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.

[24] M. Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. Emotions in argumentation: an empirical evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, 2015*, pages 156–163. AAAI Press, 2015.

[25] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[26] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations, 2020.

[27] Abeba Birhane. The unseen black faces of ai algorithms. *Nature*, 610:451–452, 2022.

[28] Abeba Birhane, Vinay Uday Prabhu, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. Into the laion's den: Investigating hate in multimodal datasets. In *NeurIPS*, 2023.

[29] Lynn Bliss and Allyssa Mccabe. Comparison of discourse genres: Clinical implications. *Contemporary Issues in Communication Science and Disorders*, 33, 10 2006.

[30] Daniela Braga and Maria Aldina Marques. The pragmatics of prosodic features in the political debate. In *Proc. Speech Prosody 2004*, pages 321–324, 2004.

[31] Hervé Bredin. pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Proc. Interspeech 2017*, pages 3587–3591, 2017.

[32] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, pages 1983–1987, 2023.

[33] Giovanni Briganti and JérÃt̂me R. Lechien. Speech and voice quality as digital biomarkers in depression: A systematic review. *Journal of Voice*, 2025.

[34] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127. Association for Computational Linguistics, July 2018.

[35] Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 5427–5433, 2018.

[36] Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, Rui Li, Zhihua Guo, Rong Ma, Jing Yang, Lan Zhang, Xiping Hu, Yumin Li, and Bin Hu. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178, Apr 2022.

[37] Amparo Elizabeth Cano-Basave and Yulan He. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California, June 2016. Association for Computational Linguistics.

[38] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training. *arXiv preprint*, arXiv:1810.06583, 2020.

[39] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091, 2022.

[40] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thummler, Esma ISMAILOVA, Nicholas W., Francois Bremond, Massimiliano Todisco, Maria A Zuluaga, and Laura M. Ferrari. StressID: a multimodal dataset for stress identification. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[41] Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

[42] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321âĂŞ357, June 2002.

[43] Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States, July 2022. Association for Computational Linguistics.

[44] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*

32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 8928–8939, 2019.

[45] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[46] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[47] Shuaiqi Chen, Xiaofen Xing, Guodong Liang, and Xiangmin Xu. I feel stressed out: A mandarin speech stress dataset with new paradigm. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 583–589, 2022.

[48] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.*, 13:359–393, 1999.

[49] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785âĂŞ794, New York, NY, USA, 2016. Association for Computing Machinery.

[50] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[51] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.*, 2(12):772–782, 2020.

[52] Subrahmanyasarma Chitta, Justin Crawly, Sai Ganesh Reddy, and Dheeraj Kumar. Balancing data sharing and patient privacy in interoperable health systems. *Distributed Learning and Broad Applications in Scientific Research*, 5:886–925, Oct. 2019.

[53] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, and Stefan Scherer. A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 488–495, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[54] Shreyan Chowdhury, Verena Praher, and Gerhard Widmer. Tracing back music emotion predictions to sound sources and intuitive perceptual qualities, 2021.

[55] Elvan Ciftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6, 2018.

[56] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[57] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[58] Joana Correia, Bhiksha Raj, Isabel Trancoso, and Francisco Teixeira. Mining Multimodal Repositories for Speech Affecting Diseases. In *Proc. Interspeech 2018*, pages 2963–2967, 2018.

[59] Albin Correya, Romain Hennequin, and Mickaël Arcos. Large-scale cover song detection in digital music libraries using metadata, lyrics and audio features. *arXiv preprint arXiv:1808.10351*, 2018.

[60] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint*, arXiv:2005.11262, 2020.

[61] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November 2020. Association for Computational Linguistics.

[62] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[63] Yinglong Dai, Zheng Yan, Jiangchang Cheng, Xiaojun Duan, and Guojun Wang. Analysis of multimodal data fusion from an information theory perspective. *Information Sciences*, 623:164–183, 2023.

[64] Klara Daly and Oluwafemi Olukoya. Depression detection in read and spontaneous speech: A multimodal approach for lesser-resourced languages. *Biomedical Signal Processing and Control*, 108:107959, 2025.

[65] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[66] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, arXiv:2501.12948, 2025.

[67] Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[68] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. Assessing language model deployment with risk cards. *arXiv preprint*, arXiv:2303.18190, 2023.

[69] Francesca D'Errico, Rosario Signorello, Didier Demolin, and Isabella Poggi. The perception of charisma from voice: A cross-cultural study. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 552–557, 09 2013.

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[71] Mark Díaz, Ian D. Kivlichan, Rachel Rosen, Dylan K. Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2342–2351. ACM, 2022.

[72] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE J. Biomed. Health Inform.*, 22(2):525–536, March 2018.

[73] Victor Dibia. NeuralQA: A usable library for question answering (contextual query expansion + BERT) on large datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 15–22, Online, October 2020. Association for Computational Linguistics.

[74] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[75] Ada Dogrucu, Alex Perucic, Anabella Isaro, Damon Ball, Ermal Toto, Elke A. Rundensteiner, Emmanuel Agu, Rachel Davis-Martin, and Edwin Boudreaux. Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health*, 17:100118, 2020.

[76] Xingjian Du, Ke Chen, Zijie Wang, Bilei Zhu, and Zejun Ma. ByteCover2: Towards dimensionality reduction of latent embedding for efficient cover song identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 616–620, 2022.

[77] Xingjian Du, Mingyu Liu, Pei Zou, Xia Liang, Zijie Wang, Huidong Liang, and Bilei Zhu. X-cover: Better music version identification system by integrating pretrained ASR model. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 70–77, 2024.

[78] Xingjian Du, Zijie Wang, Xia Liang, Huidong Liang, Bilei Zhu, and Zejun Ma. ByteCover3: Accurate cover song identification on short queries. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[79] Tatjana Dubovičienė and Pavel Skorupa. The analysis of some stylistic features of english advertising slogans. *Žmogus ir Žodis*, 16:61–75, 2014.

[80] Malikeh Ehghaghi, Frank Rudzicz, and Jekaterina Novikova. Data-driven approach to differentiating between depression and dementia from noisy speech and language data. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 24–37, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics.

[81] Nour El Houda Dehimi and Zakaria Tolba. Attention mechanisms in deep learning : Towards explainable artificial intelligence. In *2024 6th Inter-*

national Conference on Pattern Analysis and Intelligent Systems (PAIS), pages 1–7, 2024.

[82] Catherine K Ettman and Sandro Galea. The potential influence of AI on population mental health. *JMIR Ment. Health*, 10:e49936, November 2023.

[83] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[84] Marta Fadda, Martin Sykora, Suzanne Elayan, Milo A Puhan, John A Naslund, Stephen J Mooney, Emiliano Albanese, Rosalba Morese, and Oliver Gruebner. Ethical issues of collecting, storing, and analyzing geo-referenced tweets for mental health research. *Digit. Health*, 8:205520762210925, January 2022.

[85] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1):78–88, April 2021.

[86] William Falcon and The PyTorch Lightning team. PyTorchLightning, March 2019.

[87] Anna Favaro, Yi-Ting Tsai, Ankur Butala, Thomas Thebaud, Jesús Villalba, Najim Dehak, and Laureano Moro-Velázquez. Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, 166:107559, 2023.

[88] Tiantian Feng, Rajat Hebbar, Nicholas Mehlman, Xuan Shi, Aditya Kommineni, and Shrikanth Narayanan. A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness. *arXiv preprint*, arXiv:2212.09006, 2022.

[89] Margaret Forbes, Davida Fromm, and Brian Macwhinney. Aphasiabank: A resource for clinicians. *Seminars in speech and language*, 33:217–22, 08 2012.

[90] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021.

[91] M. Gheorghe, S. Mihalache, and D. Burileanu. Using deep neural networks for detecting depression from speech. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 411–415, Helsinki, Finland, 2023.

[92] John Gideon, Emily Mower Provost, and Melvin McInnis. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, March 2016.

[93] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 80–89. IEEE, 2018.

[94] Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[95] Piers Gooding and Timothy Kariotis. Ethics and law in research on algorithmic and data-driven technology in mental health care: Scoping review. *JMIR Ment Health*, 8(6):e24668, Jun 2021.

[96] Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, December 2025.

[97] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David Devault, Stacy Marsella, David Traum, Albert Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[98] Leo Groarke and Gabrijela Kišiček. Sound arguments: An introduction to auditory argument. In *Argumentation and inference: Proceedings of 2nd European Conference on Argumentation*, pages 177–198. London: Collage Publications, 2018.

[99] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL (1)*, pages 2225–2235. Association for Computational Linguistics, 2018.

[100] Weitong Guo, Hongwu Yang, Zhenyu Liu, Yaping Xu, and Bin Hu. Deep neural networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks. *Front. Neurosci.*, 15, April 2021.

[101] Shubham Gupta, Mirco Ravanelli, Pascal Germain, and Cem Subakan. Phoneme discretized saliency maps for explainable detection of ai-generated voice. In *Interspeech 2024*, pages 3295–3299, 2024.

[102] G. Gutierrez, C. Stephenson, J. Eadie, K. Asadpour, and N. Alavi. Examining the role of ai technology in online mental healthcare: Opportunities, challenges, and implications, a mixed-methods review. *Frontiers in Psychiatry*, 15:1356773, 2024.

[103] Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy, July 2019. Association for Computational Linguistics.

[104] Charles Leonard Hamblin. Fallacies. In *Advanced Reasoning Forum*. Advanced Reasoning Forum, 2022.

[105] John H. L. Hansen and Sahar E. Bou-Ghazale. Getting started with SUSAS: a speech under simulated and actual stress database. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 1743–1746, 1997.

[106] Sahar Harati, Andrea Crowell, Yijian Huang, Helen Mayberg, and Shamim Nemati. Classifying depression severity in recovery from major depressive

disorder via dynamic facial features. *IEEE Journal of Biomedical and Health Informatics*, 24(3):815–824, 2020.

[107] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[108] Verena Haunschmid, Ethan Manilow, and Gerhard Widmer. audiolime: Listenable explanations using source separation. *arXiv preprint*, arXiv:2008.00582, 2020.

[109] M. He, E. M. Bakker, and M. S. Lew. Dpd (depression detection) net: a deep neural network for multimodal depression detection. *Health Information Science and Systems*, 12(1):53, Nov 2024.

[110] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023.

[111] Chadi Helwe, Chloé Clavel, and Fabian Suchanek. LogiTorch: A PyTorch-based library for logical reasoning on natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–257, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics.

[112] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 31–35. IEEE, 2016.

[113] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970.

[114] Andre Holzapfel, Bob L. Sturm, and Mark Coeckelbergh. Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, 1(1):44–55, 2018.

[115] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks, 2019.

[116] Oleh Hornykiewicz. Biochemical aspects of parkinson's disease. *Neurology*, 51:S2–S9, 1998.

[117] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 2021.

[118] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, 74(2):171, 2019.

[119] Andreea Iana, Goran Glavaš, and Heiko Paulheim. NewsRecLib: A PyTorch-lightning library for neural news recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 296–310, Singapore, December 2023. Association for Computational Linguistics.

[120] Keith Ito and Linda Johnson. The LJ speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[121] Petar Ivanov, Ivan Koychev, Momchil Hardalov, and Preslav Nakov. Detecting check-worthy claims in political debates, speeches, and interviews using audio data. *arXiv preprint*, arXiv:2306.05535, 2023.

[122] Olga Ivanova, Juan José G. Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E. Llorente, and Nuria Carcavilla González. Discriminating speech traits of alzheimer's disease assessed through a corpus of reading task for spanish language. *Comput. Speech Lang.*, 73(C), may 2022.

[123] Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. MuSE: a multimodal dataset of stressed emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1499–1510, Marseille, France, May 2020. European Language Resources Association.

[124] Xiaowen Jia, Jingxia Chen, Kexin Liu, Qian Wang, and Jialing He. Multimodal depression detection based on an attention graph convolution and transformer. *College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology*, 2025.

[125] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[126] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[127] A-Hyeon Jo and Keun-Chang Kwak. Diagnosis of depression based on four-stream model of bi-lstm and cnn from audio and text information. *IEEE Access*, 10:134113–134135, 2022.

[128] David S. Johnson, Olya Hakobyan, Jonas Paletschek, and Hanna Drimalla. Explainable ai for audio and visual affective computing: A scoping review. *IEEE Transactions on Affective Computing*, 16(2):518–536, 2025.

[129] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502, 2004.

[130] Akhil P. Joseph and Anithamol Babu. The unseen dilemma of ai in mental healthcare. *AI & SOCIETY*, 2024(5):13, 2024.

[131] Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.

[132] Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E. Morgan. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. *Psychiatry Research*, 352:116690, 2025.

[133] Warale Avinash Kalyan, Siddharth Pagaria, Chaitra V, and Spoorthi H G. Multimodal argumentative fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*, 2025.

[134] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

[135] Michelle R Kaufman, Evan L Eschliman, and Tahilin Sanchez Karver. Differentiating sex and gender in health research to achieve gender equity. *Bulletin of the World Health Organization*, 101(10):666, 2023.

[136] Hassan Salami Kavaki and Michael I. Mandel. Identifying Important Time-Frequency Locations in Continuous Speech Utterances. In *Proc. Interspeech 2020*, pages 1639–1643, 2020.

[137] Daniel Kempler, Susan Curtiss, and Catherine Jackson. Syntactic preservation in alzheimer's disease. *Journal of speech and hearing research*, 30:343–50, 10 1987.

[138] Bangul Khan, Hajira Fatima, Ayatullah Qureshi, Sanjay Kumar, Abdul Hanan, Jawad Hussain, and Saad Abdullah. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials & Devices*, pages 1–8, February 2023. Epub ahead of print.

[139] S. Khan, S. M. Umar Saeed, J. Frnda, A. Arsalan, R. Amin, R. Gantassi, and S. H. Noorani. A machine learning based depression screening framework using temporal domain features of the electroencephalography signals. *PLoS ONE*, 19(3):e0299127, Mar 27 2024.

[140] Hyeoneui Kim, Elizabeth Bell, Jihoon Kim, Amy Sitapati, Joe Ramsdell, Claudiu Farcas, Dexter Friedman, Stephanie Feudjio Feupe, and Lucila Ohno-Machado. iCONCUR: informed consent for clinical data and biosample use for research. *J. Am. Med. Inform. Assoc.*, 24(2):380–387, March 2017.

[141] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. Do stakeholder needs differ? Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. *International Journal of Human-Computer Studies*, 181:103160, 2024.

[142] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[143] Gabrijela Kišiček. The role of prosodic features in the analysis of multimodal argumentation. In *International Society for the Study of Argumentation (ISSA), 8th international conference on argumentation*. Rozenberg Quarterly, The Magazine, 2014.

[144] Gabrijela Kišiček. Listen carefully! fallacious auditory arguments. In *Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*, OSSA 12, pages 17–32. University of Windsor, 2020.

[145] Gabrijela Kišiček. Listen carefully! fallacious auditory arguments. In *Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*, OSSA 12, pages 17–32. University of Windsor, 2020.

[146] Marius Klug, T. Berg, and Klaus Gramann. Optimizing EEG ICA decomposition with data cleaning in stationary and mobile experiments. *Scientific Reports*, 14, 06 2024.

[147] Peter Knees, Markus Schedl, and Gerhard Widmer. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 564–569, 2005.

[148] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[149] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition, 2020.

[150] Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. Detecting deception in political debates using acoustic and textual features. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 652–659, 2019.

[151] Maria Koutsombogera and Harris Papageorgiou. Multimodal indicators of persuasion in political interviews. In Isabella Poggi, Francesca D'Errico, Laura Vincze, and Alessandro Vinciarelli, editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 16–29, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[152] Kurt Kroenke and Robert L Spitzer. The PHQ-9: a new depression diagnostic and severity measure, 2002.

[153] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1):163–173, 2009.

[154] Anna Kruspe. More than words: Advancements and challenges in speech recognition for singing, 2024.

[155] Sonal Kumar, Prem Seetharaman, Justin Salamon, Dinesh Manocha, and Oriol Nieto. SILA: Signal-to-language augmentation for enhanced control in text-to-audio generation, 2024.

[156] Eli Kurtz, Youxiang Zhu, Tiffany Driesse, Bang Tran, John A. Batsis, Robert M. Roth, and Xiaohui Liang. Early detection of cognitive decline using voice assistant commands. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[157] Moreno La Quatra, Maria Francesca Turco, Torbjørn Svendsen, Giampiero Salvi, Juan Rafael Orozco-Arroyave, and Sabato Marco Siniscalchi. Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions. In *Interspeech 2024*, 2024.

[158] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J. Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8, 1999.

[159] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[160] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI

and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473, 2021.

[161] Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *Am. J. Speech. Lang. Pathol.*, 32(2):426–438, March 2023.

[162] Eva Cantín Larumbe and Adriana Chust Vendrell. Argumentative fallacy detection in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*, 2025.

[163] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, December 2019.

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

[165] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, July 2023. Association for Computational Linguistics.

[166] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3530–3537. AAAI Press, 2018.

[167] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, 134(3):103, 2026.

[168] Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*,

pages 3053–3066, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[169] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 3799–3802, New York, NY, USA, 2021. Association for Computing Machinery.

[170] Yiling Li, Yi Lin, Hongwei Ding, and Chunbo Li. Speech databases for mental disorders: A systematic review. *General Psychiatry*, 32(3), 2019.

[171] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. Conversational emotion analysis via attention mechanisms. In *INTERSPEECH*, pages 1936–1940. ISCA, 2019.

[172] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards user-centric visualizations and interpretations of multimodal models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023*, pages 214:1–214:21. ACM, 2023.

[173] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

[174] Gionnieve Lim and Simon T. Perrault. Evaluation of an llm in identifying logical fallacies: A call for rigor when adopting LLMs in HCI research. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, pages 303–308, New York, NY, USA, 2024. Association for Computing Machinery.

[175] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr DollÃąr. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[176] Weizhe Lin, Indigo Orton, Qingbiao Li, Gabriela Pavarini, and Marwa Mahmoud. Looking at the body: Automatic analysis of body gestures and self-

adaptors in psychological distress. *IEEE Transactions on Affective Computing*, 14(2):1175–1187, 2023.

[177] Marco Lippi and Paolo Torroni. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2979–2985. AAAI Press, 2016.

[178] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2), mar 2016.

[179] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, arXiv:1907.11692, 2019.

[180] Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, pages 120–132, Singapore, December 2023. Association for Computational Linguistics.

[181] Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. ImageArg: A multimodal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.

[182] Pavel Llamocca, Victoria López, Matilde Santos, and Milena Čukić. Personalized characterization of emotional states in patients with bipolar disorder. *Mathematics*, 9(11), 2021.

[183] Romain Loiseau, Baptiste Bouvier, Yann Teytaut, Elliot Vincent, Mathieu Aubry, and Loïc Landrieu. A model you can hear: Audio identification with playable prototypes. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, pages 694–700, 2022.

[184] Clara Lombardo, Giulia Esposito, Silvia Carbone, Salvatore Serrano, and Carmela Mento. Speech analysis and speech emotion recognition in mental disease: a scoping review. *Frontiers in Psychology*, 16, 2025.

[185] Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.

[186] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[187] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3, 2021.

[188] Marta Maffia, Loredana Schettino, and Vincenzo Norman Vitale. Automatic detection of parkinson's disease with connected speech acoustic features: towards a linguistically interpretable approach. In *Proceedings of the 9th Italian Conference on Computational Linguistics CLiC-it 2023: Venice, Italy, November 30-December 2, 2023*. Accademia University Press, 2024.

[189] Vidula Magdum, Omkar Jayant Dhekane, Sharayu Sandeep Hiwarkhedkar, Saloni Sunil Mittal, and Raviraj Joshi. mahaNLP: A Marathi natural language processing library. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 34–40, Bali, Indonesia, November 2023. Association for Computational Linguistics.

[190] Ian Magnusson, Noah A. Smith, and Jesse Dodge. Reproducibility in NLP: what have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12789–12811. Association for Computational Linguistics, 2023.

[191] Khyati Mahajan and Samira Shaikh. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online, July 2021. Association for Computational Linguistics.

[192] Muhammad Shairoze Malik. Analyzing the trade-offs of data sharing in

social networks and privacy concerns. *International Journal for Electronic Crime Investigation*, 8(4), 2024.

[193] Eleonora Mancini, Andrea Galassi, Federico Ruggeri, and Paolo Torroni. Disruptive situation detection on public transport through speech emotion recognition. *Intelligent Systems with Applications*, 21:200305, 2024.

[194] Eleonora Mancini, Francesco Paissan, Mirco Ravanelli, and Cem Subakan. Lmac-td: Producing time domain explanations for audio classifiers. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. DOI: 10.1109/ICASSP49660.2025.10890448.

[195] Eleonora Mancini, Francesco Paissan, Paolo Torroni, Mirco Ravanelli, and Cem Subakan. Investigating the effectiveness of explainability methods in parkinson's detection from speech. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2025. DOI: 10.1109/ICASSPW65056.2025.11011035.

[196] Eleonora Mancini, Federico Ruggeri, Stefano Colamonaco, Andrea Zecca, Samuele Marro, and Paolo Torroni. MAMKit: A comprehensive multimodal argument mining toolkit. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 69–82, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[197] Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.

[198] Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[199] Eleonora Mancini, Federico Ruggeri, Serena Villata, and Paolo Torroni. Overview of mm-argfallacy2025 on multimodal argumentative fallacy de-

tection and classification in political debates. In *Proceedings of the 12th Argument mining Workshop*, pages 358–368, 2025.

[200] Eleonora Mancini, Joan Serrà, Paolo Torroni, and Yuki Mitsufuji. Leveraging whisper embeddings for audio-based lyrics matching. *arXiv preprint*, arXiv:2510.08176, 2025.

[201] Eleonora Mancini, Ana Tanevska, Andrea Galassi, Alessio Galatolo, Federico Ruggeri, and Paolo Torroni. Promoting the responsible development of speech datasets for mental health and neurological disorders research. *J. Artif. Intell. Res.*, 82:937–972, 2025.

[202] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics, 2014.

[203] Nicole Martinez-Martin, Henry T Greely, and Mildred K Cho. Ethical development of digital phenotyping tools for mental health applications: Delphi study. *JMIR Mhealth Uhealth*, 9(7):e27343, Jul 2021.

[204] C. D. Mathers and D. Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLOS Medicine*, 3:1–20, November 2006.

[205] Melkamu Mersha, Khang Lam, Joseph Wood, Ali K. AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, 2024.

[206] Rafael Mestre, Stuart Middleton, Matt Ryan, Masood Gheasi, Timothy Norman, and Jiatong Zhu. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 274–288, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[207] Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. M-arg: Multimodal argument mining dataset for

political debates with audio and transcripts. In *ArgMining@EMNLP*, pages 78–88. Association for Computational Linguistics, 2021.

[208] Microsoft. Large-scale self-supervised pre-training across tasks, languages, and modalities., 2020.

[209] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[210] Adam S. Miner, Albert Haque, Jason A. Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A. Arnow, W. Stewart Agras, Li Fei-Fei, and Nigam H. Shah. Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*, 3(1):82, 2020.

[211] Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. A recorded debating dataset. In *LREC*. European Language Resources Association (ELRA), 2018.

[212] Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[213] Saumitra Mishra, Emmanouil Benetos, Bob L. Sturm, and Simon Dixon. Reliable local explanations for machine listening. *arXiv preprint*, arXiv:2005.07788, 2020.

[214] Saumitra Mishra, Bob L. Sturm, and Simon Dixon. Local interpretable model-agnostic explanations for music content analysis. In *International Society for Music Information Retrieval Conference*, 2017.

[215] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and

Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM, 2019.

[216] Rachel L.C. Mitchell and Elliott D. Ross. Attitudinal prosody: What we know and directions for future study. *Neuroscience and Biobehavioral Reviews*, 37(3):471–479, 2013.

[217] Saif M. Mohammad. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8368–8379. Association for Computational Linguistics, 2022.

[218] Christoph Molnar. *Interpretable Machine Learning*. Online, 2 edition, 2022.

[219] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online, October 2020. Association for Computational Linguistics.

[220] Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai-Doss, and Sébastien Marcel. Understanding and Visualizing Raw Waveform-Based CNNs. In *Proc. Interspeech 2019*, pages 2345–2349, 2019.

[221] Charlene L Muehlenhard and Zoe D Peterson. Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex roles*, 64:791–803, 2011.

[222] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020.

[223] Wajid Mumtaz and Abdul Qayyum. A deep learning framework for automatic diagnosis of unipolar depression. *International Journal of Medical Informatics*, 132:103983, 2019.

[224] Habibeh Naderi, Behrouz Haji Soleimani, Sheri Rempel, Stan Matwin, and Rudolf Uher. Multimodal deep learning for mental disorders prediction from audio speech samples. *arXiv preprint*, arXiv:1909.01067, 2019.

[225] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–387, Cham, 2018. Springer International Publishing.

[226] Siddharth Narayanaswamy, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank D. Wood, and Philip H. S. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5925–5935, 2017.

[227] NP Narendra, Björn Schuller, and Paavo Alku. The detection of parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936, 2021.

[228] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s), July 2023.

[229] Jennifer J. Newson and Tara C. Thiagarajan. Eeg frequency bands in psychiatric disorders: A review of resting state studies. *Frontiers in Human Neuroscience*, Volume 12 - 2018, 2019.

[230] Ernst Niedermeyer and Fernando Lopes da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Lippincott Williams & Wilkins, 5th edition, 2004.

[231] Jekaterina Novikova and Aparna Balagopalan. On speech datasets in machine learning for healthcare, 01 2020.

[232] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems-an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[233] Annisaa Fitri Nurfidausi, Eleonora Mancini, and Paolo Torroni. TRI-DEP: A trimodal comparative study for depression detection using speech, text, and eeg. *arXiv preprint*, arXiv:2510.08176, 2025.

[234] M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova. Multimodal data fusion for depression detection approach. *Computation*, 13(1):9, 2025.

[235] OpenAI. Gpt-5.2 system card. Technical report, OpenAI, December 2025. Update to GPT-5 System Card.

[236] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Pearson Education, 3rd edition, 2014.

[237] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia González-Rátiva, and Elmar Nöth. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 342–347, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[238] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Lrec*, pages 342–347, 2014.

[239] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[240] Frederik Pahde, Mihai Marian Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2643–2652. IEEE, 2021.

[241] Francesco Paissan, Luca Della Libera, Mirco Ravanelli, and Cem Subakan. Listenable maps for zero-shot audio classifiers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[242] Francesco Paissan, Mirco Ravanelli, and Cem Subakan. Listenable maps for audio classifiers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 39009–39021, 2024.

[243] Francesco Paissan, Cem Subakan, and Mirco Ravanelli. Posthoc interpretation via quantization. *arXiv preprint*, arXiv:2303.12659, 2023.

[244] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Theodore Alexander Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 881–904. ACM, 2023.

[245] Jayneel Parekh, Sanjeel Parekh, Pavlo Mozharovskyi, Florence d'Alché-Buc, and Gaël Richard. Listen to interpret: Post-hoc interpretability for audio networks with nmf. In *Advances in Neural Information Processing Systems*, volume 35, pages 35270–35283, 2022.

[246] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. Retrieving similar lyrics for music recommendation system. In *Proceedings of the International Conference on Natural Language Processing (ICON)*, pages 290–297, 2017.

[247] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[248] Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. How susceptible are LLMs to logical fallacies? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8276–8286, Torino, Italia, May 2024. ELRA and ICCL.

[249] Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[250] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[251] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.

[252] Serge Pinto, Canan Ozsancak, Elina Tripoliti, Stéphane Thobois, Patricia Limousin-Dowsey, and Pascal Auzou. Treatments for dysarthria in parkinson's disease. *The Lancet Neurology*, 3(9):547–556, 2004.

[253] Alessio Pittiglio. Leveraging context for multimodal fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*, 2025.

[254] Werner Poewe, Klaus Seppi, Caroline M Tanner, Glenda M Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E Lang. Parkinson disease. *Nature reviews Disease primers*, 3(1):1–21, 2017.

[255] Claire Polo, Kristine Lund, Christian Plantin, and Gerald P. Niccolai. Group emotions: the social and cognitive functions of emotions in argumentation. *Int. J. Comput. Support. Collab. Learn.*, 11(2):123–156, 2016.

[256] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1776–1826. ACM, 2022.

[257] PyTorch. Torchmultimodal (beta release), 2022.

[258] A. Qayyum, I. Razzak, and W. Mumtaz. Hybrid deep shallow network for assessment of depression using electroencephalogram signals. In *International Conference on Neural Information Processing*, pages 245–257. Springer, Cham, 2020.

[259] A. Qayyum, I. Razzak, M. Tanveer, M. Mazhar, and B. Alhaqbani. High-density electroencephalography and speech signal-based deep framework for clinical depression diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20:2587–2597, 2023.

[260] Kun Qian, Ruolan Huang, Zhihao Bao, Yang Tan, Zhonghao Zhao, Mengkai Sun, Bin Hu, Björn W. Schuller, and Yoshiharu Yamamoto. Detecting somatisation disorder via speech: introducing the shenzhen somatisation speech corpus, April 2023.

[261] Changqin Quan, Kang Ren, Zhiwei Luo, Zhonglue Chen, and Yun Ling. End-to-end deep learning approach for parkinson's disease detection from speech signals. *Biocybernetics and Biomedical Engineering*, 42(2):556–574, 2022.

[262] Lawrence Rabiner and Ronald Schafer. *Theory and Applications of Digital Speech Processing*. Prentice Hall Press, USA, 1st edition, 2010.

[263] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.

[264] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[265] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518, 2023.

[266] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *Online*, 2018.

[267] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[268] Dhanesh Ramachandram and Graham W. Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[269] Vikram Ramanarayanan. Multimodal technologies for remote assessment of neurological and mental health. *Journal of Speech, Language, and Hearing Research*, 67(11):4233–4245, November 2024. Epub 2024 Jul 10.

[270] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Int. Res.*, 73, may 2022.

[271] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. Open-source conversational AI with SpeechBrain 1.0. *arXiv preprint*, arXiv:2407.00463, 2024.

[272] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement. In *Proc. Interspeech 2022*, pages 3338–3342, 2022.

[273] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88, 2019.

[274] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[275] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107, 2022.

[276] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016.

[277] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Summary for avec 2018: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 2111–2112, New York, NY, USA, 2018. Association for Computing Machinery.

[278] Cristian David Rios-Urrego, Santiago Andres Moreno-Acevedo, Elmar Nöth, and Juan Rafael Orozco-Arroyave. End-to-end parkinson's disease detection using a deep convolutional recurrent network. In *International Conference on Text, Speech, and Dialogue*, pages 326–338. Springer, 2022.

[279] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine A. Heller. Healthsheet: Development of a transparency artifact for health datasets. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1943–1961. ACM, 2022.

[280] Giovanni Rubeis. ihealth: The ethics of artificial intelligence and big data in mental healthcare. *Internet Interv.*, 28(100518):100518, April 2022.

[281] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[282] Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. VivesDebate-speech: A corpus of spoken argumentation to leverage audio features for argument mining. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2071–2077, Singapore, December 2023. Association for Computational Linguistics.

[283] Jan Rusz, Roman Cmejla, Tereza Tykalova, Hana Ruzickova, Jiri Klempir, Veronika Majerova, Jana Picmausova, Jan Roth, and Evzen Ruzicka. Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3):2171–2181, 2013.

[284] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE J Biomed Health Inform*, 17(4):828–834, July 2013.

[285] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986.

[286] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[287] D. Santomauro, A. Mantilla Herrera, J. Shadid, and P. Zheng. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398, October 2021.

[288] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics.

[289] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*, pages 3465–3469. ISCA, 2019.

[290] M. Schoeffler and et al. webmushra - a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 2018.

[291] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[292] Bjorn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley Publishing, 1st edition, 2013.

[293] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[294] Joan Serrà, R. Oguz Araz, Dmitry Bogdanov, and Yuki Mitsufuji. Supervised contrastive learning from weakly-labeled audio segments for musical version matching. In *International Conference on Machine Learning (ICML)*, 2025.

[295] Khaled Shaalan. A survey of arabic named entity recognition and classification. *Comput. Linguistics*, 40(2):469–510, 2014.

[296] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[297] Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 850–861. ACM, 2021.

[298] Ying Shen, Huiyu Yang, and Lin Lin. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251, 2022.

[299] Jiatong Shi, Shuai Guo, Tao Qian, Tomoki Hayashi, Yuning Wu, Fangzheng Xu, Xuankai Chang, Huazhe Li, Peter Wu, Shinji Watanabe, and Qin Jin. Muskits: an End-to-end Music Processing Toolkit for Singing Voice Synthesis. In *Proc. Interspeech 2022*, pages 4277–4281, 2022.

[300] Anastasia Shimorina and Anya Belz. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems*

(HumEval), pages 54–75, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[301] Rosario Signorello. *The Vocal Attractiveness of Charismatic Leaders*, pages 41–54. Springer Singapore, Singapore, 2021.

[302] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[303] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

[304] Pavel Skorupa and Tatjana Dubovičienė. Linguistic characteristics of commercial and social advertising slogans. *Coactivity: Philology, Educology*, 23:108–118, 12 2015.

[305] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint*, arXiv:1706.03825, 2017.

[306] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2951âĂŞ2959, Red Hook, NY, USA, 2012. Curran Associates Inc.

[307] Kacper Sokol and Peter A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 56–67. ACM, 2020.

[308] Bhakti Sonawane and Priyanka Sharma. Speech-based solution to parkinson's disease management. *Multimedia Tools and Applications*, 80(19):29437–29451, 2021.

[309] Chenguang Song, Ke Chao, Bingjing Jia, and Yiqing Shen. Multimodal prototypical network for interpretable sentiment classification. *Scientific Reports*, 15(1):32939, 2025.

[310] Meishu Song, Andreas Triantafyllopoulos, Zijiang Yang, Hiroki Takeuchi, Toru Nakamura, Akifumi Kishi, Tetsuro Ishizawa, Kazuhiro Yoshiuchi, Xin Jing, Vincent Karas, Zhonghao Zhao, Kun Qian, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. Daily mental health monitoring from speech: A real-world japanese dataset and multitask learning analysis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[311] Olaf Sporns. Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23(2):162–171, April 2013.

[312] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[313] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[314] Igor Stancin, Mario Cifrek, and Alan Jovic. A review of eeg signal features and their application in driver drowsiness detection systems. *Sensors*, 21:3786, 05 2021.

[315] Felix Stollenwerk. nerblackbox: A high-level library for named entity recognition in python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 174–178, Singapore, December 2023. Association for Computational Linguistics.

[316] Julia Stoyanovich and Bill Howe. Nutritional labels for data and models. *IEEE Data Eng. Bull.*, 42(3):13–23, 2019.

[317] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 21–25. IEEE, 2021.

[318] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and

Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP*, 2021.

[319] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017.

[320] Abdullah Tahir, Imaan Ibrar, Huma Ameer, Mehwish Fatima, and Seemab Latif. Prompt-guided augmentation and multi-modal fusion for argumentative fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*, 2025.

[321] Tianyi Tang, Junyi Li, Zhipeng Chen, Yiwen Hu, Zhuohao Yu, Wenxun Dai, Wayne Xin Zhao, Jian-yun Nie, and Ji-rong Wen. TextBox 2.0: A text generation library with pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 435–444, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics.

[322] Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection. In *Proc. INTERSPEECH 2023*, pages 4149–4153, 2023.

[323] Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. DEPAC: a corpus for depression and anxiety detection from speech. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–16, Seattle, USA, July 2022. Association for Computational Linguistics.

[324] Yla Tausczik and James Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54, 03 2010.

[325] Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, USA, 1st edition, 2009.

[326] TencentGameMate. Chinese HuBERT Large, 2024.

[327] Mohamed Terra, Mohamed Baklola, Shaimaa Ali, and Karim El-Bastawisy. Opportunities, applications, challenges and ethical implications of artificial

intelligence in psychiatry: a narrative review. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, 59(1):80, Jun 2023.

[328] A. Thakkar, A. Gupta, and A. De Sousa. Artificial intelligence in positive mental health: a narrative review. *Frontiers in Digital Health*, 6:1280235, 2024.

[329] Gyanendra Tiwary, Shivani Chauhan, and K. K. Goyal. Automatic depression detection using multi-modal & late-fusion based architecture. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 1–6, 2023.

[330] ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering, Ermal Toto, and Elke A. Rundensteiner. Studentsadd: Rapid mobile depression and suicidal ideation screening of college students during the coronavirus pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(2), jul 2022.

[331] Ml Tlachac, Ermal Toto, Joshua Lovering, Rimsha Kayastha, Nina Taurich, and Elke Rundensteiner. Emu: Early mental health uncovering framework and dataset. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1311–1318, 2021.

[332] Ermal Toto, M. L. Tlachac, and Elke A. Rundensteiner. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *CIKM*, pages 4145–4154. ACM, 2021.

[333] Wiebke Toussaint, Lauriane Gorce, and Aaron Yi Ding. Design guidelines for inclusive speaker verification evaluation datasets. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1293–1297. ISCA, 2022.

[334] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971, 2023.

[335] Viet Anh Trinh, Brian McFee, and Michael I Mandel. Bubble Cooperative Networks for Identifying Important Speech Cues. In *Proc. Interspeech 2018*, pages 1616–1620, 2018.

[336] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-Wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8479–8492. Association for Computational Linguistics, 2022.

[337] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics.

[338] U.S. Copyright Office. Copyright registration of musical compositions and sound recordings. Circular 56A, U.S. Copyright Office, 2021.

[339] Andrea Vaglio, Romain Hennequin, Manuel Moussallam, and Gael Richard. The Words Remain The Same: Cover Detection with Lyrics Transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[340] Betty van Aken, Jens-Michalis Papaioannou, Marcel Ganesh Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix A. Gers, and Alexander Löser. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 172–184. Association for Computational Linguistics, 2022.

[341] Sari M Van Anders. Beyond sexual orientation: Integrating gender/sex

and diverse sexualities via sexual configurations theory. *Archives of sexual behavior*, 44:1177–1213, 2015.

[342] S. Varlokosta, Spyridoula Stamouli, A. Karasimos, G. Markopoulos, M. Kakavoulia, M. Nerantzini, A. Pantoula, V. Fyndanis, A. Economou, and A. Protopapas. A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC-2016), Workshop âĂIJResources and Processing of linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID-2016)*, pages 14–21, Portorovz, Slovenia, May 23, 2016 2016.

[343] Daniel Varona and Juan Luis Suárez. Discrimination, bias, fairness, and trustworthy ai. *Applied Sciences*, 12(12):5826, 2022.

[344] Juan Camilo Vásquez-Correa, Cristian David Rios-Urrego, Tomas Arias-Vergara, Maria Schuster, Jan Rusz, Elmar Noeth, and Juan Rafael Orozco-Arroyave. Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. *Pattern Recognition Letters*, 150:272–279, 2021.

[345] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[346] Thomas Vaughan-Johnston, Joshua Guyer, Leandre Fabrigar, and Charlie Shen. The role of vocal affect in persuasion: The civa model. *Journal of Nonverbal Behavior*, 45, 07 2021.

[347] David Vilares and Yulan He. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[348] Serena Villata, Elena Cabrio, Imène Jraidi, M. Sahbi Benlamine, Maher Chaouachi, Claude Frasson, and Fabien Gandon. Emotions and personality traits in argumentation: An empirical evaluation. *Argument Comput.*, 8(1):61–87, 2017.

[349] J. Volkmann, S. Stevens, and E. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8:208–208, 06 2005.

[350] Douglas N. Walton. *Informal Fallacies: Towards a Theory of Argument Criticisms*. John Benjamins, Philadelphia, 1987.

[351] Bo Wang, Yue Wu, Niall Taylor, Terry Lyons, Maria Liakata, Alejo J. Nevado-Holgado, and Kate E.A. Saunders. Learning to Detect Bipolar Disorder and Borderline Personality Disorder with Language and Speech in Non-Clinical Interviews. In *Proc. Interspeech 2020*, pages 437–441, 2020.

[352] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: Afoundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

[353] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Trans. Vis. Comput. Graph.*, 28(1):802–812, 2022.

[354] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[355] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211, 2018.

[356] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *ISMIR*, pages 825–833, 2024.

[357] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10941–10950, 2020.

[358] Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[359] Eline Westerhout and Paola Monachesi. A pilot study for a corpus of dutch aphasic speech (codas). In *LREC*, pages 1648–1653. European Language Resources Association (ELRA), 2006.

[360] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. In *Proc. Interspeech*, September 2019.

[361] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics, 2019.

[362] Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12):230964, 12 2023.

[363] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint*, arXiv:2211.06687, 2022.

[364] Til Wykes, Jessica Lipshitz, and Stephen M. Schueller. Towards the design of ethical standards related to digital mental health and all its applications. *Current Treatment Options in Psychiatry*, 6(3):232–242, Sep 2019.

[365] Xin Xie, Zhoubo Li, Xiaohan Wang, ZeKun Xi, and Ningyu Zhang. LambdaKG: A library for pre-trained language model-based knowledge graph em-

beddings. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 25–33, Bali, Indonesia, November 2023. Association for Computational Linguistics.

[366] Michael Yang, Abd-Allah El-Attar, and Theodora Chaspari. Deconstructing demographic bias in speech-based machine learning models for digital health. *Frontiers in Digital Health*, 6:1351637, 2024.

[367] Wenju Yang, Jiankang Liu, Peng Cao, Rongxin Zhu, Yang Wang, Jian K. Liu, Fei Wang, and Xizhe Zhang. Attention guided learnable time-domain filterbanks for speech depression detection. *Neural Networks*, 165:135–149, 2023.

[368] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[369] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, pages 10965–10976, 2019.

[370] F. Yesiler, G. Doras, R. M. Bittner, C. J. Tralie, and J. Serrà. Audio-based musical version identification: elements and challenges. *IEEE Signal Processing Magazine*, 38(6):115–136, 2021.

[371] Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the faithfulness measurements for model interpretations. *arXiv preprint*, arXiv:2104.08782, 2021.

[372] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 11 2024.

[373] Janice D Yoder. *Women and gender: Transforming psychology.* Prentice-Hall, Inc, 1999.

[374] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12226–12234, Jun. 2022.

[375] M. Yousufi, R. Damaševičius, and R. Maskeliūnas. Multimodal fusion of eeg and audio spectrogram for major depressive disorder recognition using modified densenet121. *Brain Sciences*, 14:1018, 2024.

[376] Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7900–7913, Toronto, Canada, July 2023. Association for Computational Linguistics.

[377] Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. Learning a representation for cover song identification using convolutional neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 541–545, 2020.

[378] Yuchen Yuan, Charles Cronin, Daniel Muüllensiefen, Shinya Fujii, and Patrick E. Savage. Perceptual and automated estimates of infringement in 40 music copyright cases. *Transactions of the International Society for Music Information Retrieval*, Oct 2023.

[379] Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online, August 2021. Association for Computational Linguistics.

[380] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. Data-centric AI: perspectives and challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pages 945–948. SIAM, 2023.

[381] Guanyu Zhang, Jinghong Ma, Piu Chan, and Zheng Ye. Graph theoretical analysis of semantic fluency in patients with parkinson's disease. *Behav. Neurol.*, 2022:1–7, April 2022.

[382] Z. Zhang and J. Wang. Can ai replace psychotherapists? exploring the future of mental health care. *Frontiers in Psychiatry*, 15:1444382, 2024.

[383] Dora Zhao, Morgan Scheuerman, Pooja Chitre, Jerone Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen Pine, and Alice Xiang. A taxonomy of challenges to curating fair datasets. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

[384] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[385] Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger B. Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo. Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 343–351, 2023.

[386] Pablo Zinemanas, Martín Rocamora, Marius Miron, Frederic Font, and Xavier Serra. An interpretable deep learning model for automatic sound classification. *Electronics*, 10(7), 2021.

[387] Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838, 2023.

Appendix

A

A.1 Dataset Pre-Processing Details

In this section, we provide information on the debates that were removed owing to issues with the audio file's quality or discrepancy between the audio content and the corresponding transcript. We removed the samples corresponding to the first parliamentary debate in 1988 (Bush vs Dukakis) since the transcript is incomplete and this would have caused alignment mismatches. Regarding the two presidential debates of 2016 (Clinton vs Trump), there was no correspondence between the audio content and corresponding transcripts. Thus, we removed these debates from the original dataset.

The transcript of the first Clinton-Bush-Perot debate of 1992 has been divided into two sections by the Commission. However, the second section did not match the audio file and, thus, we removed the samples corresponding to the second section from the dataset. In the first Carter-Ford debate in 1976, the audio contains a cut of about 30 minutes. Thus, we trimmed the audio file and kept only the audio content before the cut.

A.2 Experimental Setup Details

Table 2 reports the number of samples for each cross-validation fold splits regarding the UKDebates corpus. Likewise, Table 3 provides training statistics for the M-Arg corpus. Table 4 reports the number of samples for the training, validation and test splits of MM-USElecDeb60To16. We used the following seeds for the repeated cross-validation routine: 15371, 15372, 15373. Lastly, Table 5, Table 6 and Table 7 report the class distribution for each train, validation and test split for the UKDebates, M-Arg and MM-USElecDeb60to16 corpora, respectively.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Train	246	247	247	247	247
Validation	62	62	62	62	62
Test	78	77	77	77	77

Table 2: *The number of samples for each train, validation and test fold split regarding the UKDebates corpus.*

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	1563	1563	1563	1564	1564
Validation	391	391	391	391	391
Test	489	489	489	488	488

Table 3: *The number of samples for each train, validation and test fold split regarding the M-Arg ($\gamma \geq 0.85$) corpus.*

	No. Sentences
Train	12423
Validation	6894
Test	7464

Table 4: *The number of samples for each train, validation and test split regarding the MM-USElecDeb60to16 corpus.*

	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	claim	not-claim	claim	not-claim	claim	not-claim	claim	not-claim	claim	not-claim
Train	96	150	98	149	98	149	98	149	97	150
Validation	25	37	24	38	24	38	24	38	24	38
Test	31	47	30	47	30	47	30	47	31	46

Table 5: *Class distribution for each train, validation and test split regarding the UKDebates corpus.*

	Fold 1			Fold 2			Fold 3			Fold 4			Fold 5		
	neither	attack	support	neither	attack	support	neither	attack	support	neither	attack	support	neither	attack	support
Train	1460	19	84	1460	19	84	1460	19	84	1460	20	84	1460	19	85
Validation	365	4	22	365	4	22	366	4	21	366	4	21	366	4	21
Test	457	6	26	457	6	26	456	5	27	456	5	27	456	6	26

Table 6: *Class distribution for each train, validation and test split regarding the M-Arg corpus.*

	ASD		ACC	
	arg	not-arg	claim	premise
Train	9456	2967	5029	4427
Validation	5199	1695	2814	2385
Test	5907	1557	3036	2871

Table 7: *Class distribution for each train, validation and test split regarding the MM-USElecDeb60to16 corpus.*

A.3 Model Calibration

In this section, we report the hyper-parameters set used to calibrate each described classification model. We distinguish between input configurations TA, TO, and AO. In particular, the calibration space for input configuration TA is the combination of those regarding input configurations TO and AO. Table 8 reports the

hyper-parameter set used to calibrate the BERT model. Similarly, Table 10 and 9 describe the calibration space of the SVM and Bi-LSTM baselines, respectively.

Hyper-parameter	Search Space
Input Configuration TO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Input Configuration AO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[64, 100, 128, 256, 512]$
Audio pooling	$[None, [10, 2], [5, 5], [5, 5, 5], [5], [10, 10]$
CNN filters	$[8, 64]$
CNN kernel size	$[3, 7]$

Table 8: *The hyper-parameters search space of the BERT model.*

Hyper-parameter	Search Space
Input Configuration TO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[32, 64, 128]$
Bi-LSTM layers	$[1, 2]$
GloVe embedding	$[50, 100, 200, 300]$
Learning rate	$[1e^{-3}, 1e^{-4}, 2e^{-4}]$
Input Configuration AO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[64, 100, 128, 256, 512]$
Bi-LSTM layers	$[1, 2]$
Audio pooling	$[None, [10, 2], [5, 5], [5], [5, 5, 5], [10, 10]]$

Table 9: *The hyper-parameters search space of the Bi-LSTM model.*

Hyper-parameter	Search Space
Kernel	$[rbf, linear]$
γ	$[5e^{-2}, 1e^{-2}, 1e^{-1}, 5e^{-1}, 1.]$
C	$[0.01, 0.1, 1., 10, 100]$

Table 10: *The hyper-parameters search space of the SVM model.*

A.4 Performance on Validation Splits

Table 11 reports classification performance on the validation set of the UKDebates corpus. Likewise, Table 12 and 13 report classification metrics for M-Arg and MM-USElecDeb60to16 corpora, respectively.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	**66.18**	49.86	48.00	51.58	<u>64.45</u>
M-ArgNet	**71.64**	53.07	63.25	55.02	<u>70.52</u>
glslstm	**68.80**	52.90	67.88	49.86	<u>68.06</u>
Random			37.78		

Table 11: *Average binary F1-score on the validation set of UKDebates. For each row, we report the best results in bold, second best results are underlined instead.*

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	13.26	11.54	12.75	<u>13.50</u>	**24.09**
M-ArgNet	<u>23.69</u>	20.35	23.66	13.04	**26.56**
glslstm	<u>21.83</u>	11.98	20.34	11.43	**24.62**
Random			2.62		

Table 12: *Average macro F1-score on the validation set of M-Arg ($\gamma \geq 0.85$). For each row, we report the best results in bold, second best results are underlined instead.*

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
ASD					
SVM	**68.01**	56.34	56.76	64.40	<u>67.24</u>
M-ArgNet	<u>66.71</u>	56.14	62.30	63.59	**68.53**
glslstm	68.71	58.86	<u>69.35</u>	63.01	**69.39**
Random			50.26		
ACC					
SVM	**66.17**	49.28	49.23	57.72	<u>64.34</u>
M-ArgNet	**68.48**	50.43	67.28	58.36	<u>68.01</u>
glslstm	67.78	48.27	<u>68.38</u>	58.30	**68.49**
Random			49.43		

Table 13: *Average macro F1-score on the validation set of MM-USElecDeb60to16. For each row, we report the best results in bold, second best results are underlined instead.*

B

B.1 Terminology

In this section, we provide formal definitions for the key terminology utilized throughout the Section 3.4 to facilitate clarity and understanding.

Snippet. The term *snippet* refers to the annotated fallacy within the dialogue.
Component. The term *component* denotes an argumentative element such as *claim* or *premise*.
Component Text. It signifies the text containing the component or its nearest representation within the discourse structure.
Component Label. This term indicates the type assigned to a particular argumentative component within the discourse framework.
Span. In the context of this study, *span* and *span-level annotations* refer to the characterization of fallacies by groups of words. These groups may not necessarily form complete sentences but could extend over multiple sentences or constitute substrings within a sentence.
Sentence. A *sentence* is defined as a unit of text delimited by a full stop.

B.2 External Resources for Dataset Construction

The resources provided in Mancini et al. [197] include:

- A file that establishes the correspondence between the debate IDs and the debate recordings available on the PBS NewsHour YouTube channel[1]. This file also contains timestamps necessary for removing audio content not present in the paired transcripts, such as opening and closing remarks by the moderators.

- Pre-processed transcripts that have undergone several transformations to achieve alignment between the audio files and the text. These transformations include (1) the removal of sentences in the transcripts that do not match the audio file cuts; (2) the removal of metadata (e.g., speaker information); (3) transcripts sentence splitting; (4) the extraction of transcripts corresponding to each 20-minute portion of the audio files.

[1] `https://www.youtube.com/channel/UC6ZFN9Tx6xh-skXCuRHCDpQ`

- JSON files containing the alignment timestamps for each 20-minute audio chunk in each debate. These files include the start and end timestamps of each utterance. The alignment was performed at the sentence level, such that each utterance corresponds to one sentence in the debate.

B.3 Dataset Pre-Processing Details

According to Goffredo et al. [94], only 31 out of the 39 debates in the USED corpus are annotated with fallacies. Additionally, USED-fallacy contains a new debate w.r.t. USED and MM-USED, namely the third 2016 presidential debate between Clinton and Trump. Moreover, we exclude three debates from USED-fallacy, due to discrepancies between the audio recording and the corresponding transcripts. These debates are the first 1998 parliamentary debate between Bush and Dukakis and the first two 2016 presidential debates between Clinton and Trump. For the same reason, we also exclude the second section of the first 1992 debate between Clinton, Bush, and Perot.

To simplify audio-to-text alignment, we merge the JSON alignment files provided in Mancini et al. [197] to obtain a single alignment file for each debate. We adjust the start and end timestamps of the files after the first chunk. Specifically, we add a duration equal to 20 minutes multiplied by the identifier of the chunk. For example, the timestamps of the second alignment file (chunk_id = 1) of a debate are shifted by 20 minutes.

B.4 Dataset Cleaning and Additional Alignment Operations

We notice that 10 snippets and 5 components are missing from their corresponding dialogues. Thus, we remove the corresponding samples from the corpus. Furthermore, we discover that when the first sentence of the dialogue is a duplicate sentence in the alignment file (e.g., *Ok* or *Thank you*), the timestamp that is associated with the sentence always corresponds to the last occurrence of that sentence in the alignment. We correct such misalignments manually.

B.5 Training Details

The primary focus of our work lies not in achieving absolute performance but rather in facilitating a comparative analysis across diverse modalities. In line with this objective, we determine hyperparameters grounded in our prior experience as outlined in Table 14. For all models, we employ class weights to manage training

data imbalance. Each model is trained on a single GPU (NVIDIA 2080Ti) with 12 GB dedicated memory in less than 24 hours.

Modality	Text Model	Audio Model	BS	Epochs	Seed	LR
AO	BERT/SBERT	CLAP/Wav2Vec	8	500	15371	5e-05
TA	BERT/SBERT	CLAP/Wav2Vec	8	500	15371	5e-05
TO	RoBERTa	CLAP/Wav2Vec	32	100	15371	5e-05

Table 14: *Hyper-parameters Configurations. Modality: Input Modality where AO, TA, and TO refer to audio-only, text-audio and text-only respectively, BS: Batch Size, Epochs: Number of Training Epochs, Seed: Random Seed, LR: Learning Rate.*

C

C.1 Experimental Setup Details

Model Hyper-parameters — Table 15 reports the main hyper-parameters used in our experiments. All model configurations can be inspected in the `mamkit.configs` package.

Training — Models are trained with cross-entropy loss as standard practice for classification tasks. We additionally apply class weighting to address class imbalance in all datasets except for MM-USED ACC, where weighting is not needed. We monitor validation loss during training and load the best model checkpoint based on this metric for evaluation on validation and test splits.

Hardware — We employ an NVIDIA 2080Ti GPU with 12 GB VRAM and an NVIDIA 3060Ti GPU with 8 GB VRAM to run our experiments. All experiments regarding a dataset are run on the same device for reproducibility and fair comparison. Furthermore, individual experiments were run on a single device.

General	
`optimizer`	AdamW
`batch_size`	4
`gradient accumulation steps`	3
`effective batch_size`	12
`max_epochs`	20
`early_stopping patience`	5
`early_stopping monitor`	*val_loss*
`cross-validation seeds`	42, 2024, 666
`leave-one-out seeds`	42
`train and test seeds`	42, 2024, 666

Table 15: *General model hyper-parameters in our experiments.*

C.2 Additional Results

Table 16 reports results for all model combinations evaluated.

Model	UKDebates (ASD)	M-Arg$_\gamma$ (ARC)	MM-USED (ASD)	MM-USED (ACC)	MMUSED-fallacy (AFC)
			Text Only		
BiLSTM	$.552_{\pm.047}$	$.120_{\pm.006}$	$.811_{\pm.004}$	$.663_{\pm.002}$	$.525_{\pm.113}$
BERT	$.654_{\pm.003}$	$.132_{\pm.004}$	$.824_{\pm.009}$	$.679_{\pm.004}$	$.594_{\pm.122}$
RoBERTa	$.692_{\pm.005}$	$.172_{\pm.015}$	$.839_{\pm.010}$	$.680_{\pm.001}$	$.615_{\pm.097}$
			Audio Only		
BiLSTM w/ MFCCs	$.302_{\pm.047}$	$.003_{\pm.005}$	$.722_{\pm.027}$	$.527_{\pm.004}$	$\mathbf{.657_{\pm.000}}$
BiLSTM w/ Wav2Vec2	$.376_{\pm.023}$	$.000_{\pm.000}$	$\mathbf{.774_{\pm.008}}$	$\mathbf{.596_{\pm.005}}$	$.655_{\pm.117}$
BiLSTM w/ HuBERT	$.364_{\pm.012}$	$\mathbf{.024_{\pm.012}}$	$.745_{\pm.009}$	$.566_{\pm.007}$	$.638_{\pm.000}$
BiLSTM w/ WavLM	$\mathbf{.393_{\pm.040}}$	$.010_{\pm.010}$	$.772_{\pm.015}$	$.583_{\pm.002}$	$.652_{\pm.000}$
Transformer w/ Wav2Vec2	$.440_{\pm.030}$	$.000_{\pm.000}$	$\mathbf{.771_{\pm.019}}$	$.514_{\pm.000}$	$.567_{\pm.225}$
Transformer w/ HuBERT	$.425_{\pm.033}$	$.000_{\pm.000}$	$.765_{\pm.016}$	$.524_{\pm.004}$	$\mathbf{.629_{\pm.162}}$
Transformer w/ WavLM	$\mathbf{.455_{\pm.004}}$	$.000_{\pm.000}$	$.768_{\pm.005}$	$\mathbf{.526_{\pm.004}}$	$.594_{\pm.217}$
			Text Audio		
BiLSTM w/ MFCCs	$.528_{\pm.039}$	$.065_{\pm.014}$	$.807_{\pm.010}$	$.662_{\pm.006}$	$\mathbf{.572_{\pm.099}}$
BiLSTM w/ Wav2Vec2	$\mathbf{.533_{\pm.009}}$	$.079_{\pm.014}$	$.808_{\pm.012}$	$.665_{\pm.004}$	$.505_{\pm.168}$
BiLSTM w/ HuBERT	$.409_{\pm.017}$	$.055_{\pm.020}$	$.807_{\pm.013}$	$.653_{\pm.003}$	$.456_{\pm.131}$
BiLSTM w/ WavLM	$.501_{\pm.022}$	$\mathbf{.084_{\pm.016}}$	$\mathbf{.815_{\pm.006}}$	$\mathbf{.667_{\pm.000}}$	$.526_{\pm.174}$
MM-BERT w/ Wav2Vec2	$\mathbf{662_{\pm.004}}$	$.153_{\pm.017}$	$\mathbf{841_{\pm.005}}$	$.677_{\pm.003}$	$.561_{\pm.114}$
MM-BERT w/ HuBERT	$.626_{\pm.003}$	$\mathbf{.160_{\pm.015}}$	$.840_{\pm.006}$	$.677_{\pm.004}$	$\mathbf{.599_{\pm.128}}$
MM-BERT w/ WavLM	$.654_{\pm.019}$	$.152_{\pm.008}$	$.836_{\pm.005}$	$\mathbf{.680_{\pm.004}}$	$.580_{\pm.103}$
MM-RoBERTa w/ Wav2Vec2	$.674_{\pm.009}$	$\mathbf{.178_{\pm.012}}$	$.833_{\pm.006}$	$\mathbf{.678_{\pm.003}}$	$.608_{\pm.126}$
MM-RoBERTa w/ HuBERT	$.624_{\pm.015}$	$.147_{\pm.004}$	$.837_{\pm.003}$	$.677_{\pm.008}$	$.576_{\pm.097}$
MM-RoBERTa w/ WavLM	$\mathbf{.687_{\pm.010}}$	$.165_{\pm.018}$	$\mathbf{.837_{\pm.009}}$	$\mathbf{678_{\pm.003}}$	$\mathbf{.624_{\pm.074}}$
CSA w/ Wav2Vec2	$\mathbf{.663_{\pm.014}}$	$.137_{\pm.027}$	$.822_{\pm.002}$	$\mathbf{.693_{\pm.001}}$	$.555_{\pm.118}$
CSA w/ HuBERT	$.632_{\pm.018}$	$\mathbf{.160_{\pm.015}}$	$.813_{\pm.004}$	$\mathbf{.693_{\pm.001}}$	$\mathbf{.582_{\pm.114}}$
CSA w/ WavLM	$.655_{\pm.029}$	$.155_{\pm.030}$	$\mathbf{.833_{\pm.011}}$	$.697_{\pm.001}$	$.535_{\pm.102}$
Ensemble w/ Wav2Vec2	$\mathbf{.586_{\pm.015}}$	$\mathbf{.011_{\pm.011}}$	$.825_{\pm.004}$	$\mathbf{.681_{\pm.002}}$	$\mathbf{.612_{\pm.134}}$
Ensemble w/ HuBERT	$.531_{\pm.039}$	$.010_{\pm.004}$	$.822_{\pm.007}$	$.681_{\pm.003}$	$.611_{\pm.107}$
Ensemble w/ WavLM	$.576_{\pm.006}$	$.002_{\pm.003}$	$\mathbf{.826_{\pm.011}}$	$.680_{\pm.003}$	$.605_{\pm.136}$
Mul-TA w/ Wav2Vec2	$.592_{\pm.034}$	$\mathbf{.098_{\pm.031}}$	$.826_{\pm.011}$	$.695_{\pm.001}$	$\mathbf{.605_{\pm.110}}$
Mul-TA w/ HuBERT	$\mathbf{.616_{\pm.019}}$	$.079_{\pm.053}$	$.829_{\pm.011}$	$\mathbf{.697_{\pm.003}}$	$.594_{\pm.091}$
Mul-TA w/ WavLM	$.602_{\pm.017}$	$.063_{\pm.015}$	$\mathbf{.837_{\pm.006}}$	$.690_{\pm.003}$	$.605_{\pm.082}$

Table 16: *Test classification performance on MAM datasets. In bold, the best-performing model for each configuration.*

D

D.1 Terminology

D.1.1 Target Issue

- **Depression (DP).** A mood disorder causing a persistent feeling of sadness and loss of interest.
- **Alzheimer's (AL).** A progressive neurological disorder that causes memory loss and cognitive decline.
- **Bipolar Disorder (BD).** A mental illness characterized by extreme mood swings, including episodes of mania and depression.
- **Anxiety (AX).** A feeling of worry, nervousness, or unease, typically about an imminent event or something with an uncertain outcome.
- **Stress (S).** Feelings of emotional or physical tension, often in response to challenging or demanding situations.
- **Dementia (DM).** A decline in cognitive function severe enough to interfere with daily life and activities.
- **Somatic Symptoms Disorder (SSD).** A mental health condition characterized by physical symptoms that mimic physical illness or injury, but without a clear medical cause.
- **Parkinson's (P).** A progressive nervous system disorder affecting movement, causing tremors, stiffness, and difficulty with balance and coordination.

D.1.2 Discourse Genre

- **Clinical Interviews (CI).** Conversational sessions conducted to gather information about an individual's mental health. These sessions are often either unstructured or semi-structured and may or not be conducted by a clinician. For example, in the case of crowdsourcing data collection, the sessions may not necessarily be conducted by a clinician.
- **Speech Tasks (ST).** Structured activities designed to assess various aspects of speech production, comprehension, or fluency.

- **Spontaneous Reading (SR).** Assessing reading abilities by asking individuals to read aloud a passage or text without prior preparation.
- **Spontaneous Speech (SS).** Assessing speech production and fluency by asking individuals to speak freely on a given topic without prior preparation.
- **Semantic Fluency Tests (SFT).** Assessments that measure an individual's ability to produce words belonging to a specific category within a limited time frame, often used to evaluate language and cognitive function.

D.1.3 Source of Content

- **Personal Interactions with the Participants (PI).** It refers to the direct engagement and communication between MHND professionals and participants.
- **Online Media (OM).** Audio-visual media uploaded on various online platforms, including but not limited to YouTube, or other vlogging platforms, which can encompass a wide range of content such as personal vlogs, educational material, entertainment, or any other form of audio-visual content available on the internet.
- **Crowdsourcing (CS).** The practice of obtaining information, ideas, or services by soliciting contributions from a large group of people, typically facilitated through online platforms or communities.

D.1.4 Other

- **MHND**: Mental Health and Neurological Disorders.
- **Unified Parkinson's Disease Rating Scale (UPDRS).** A comprehensive tool used to assess the severity and progression of Parkinson's disease by evaluating various aspects including motor function, activities of daily living, and mood.
- **Hoehn and Yahr Scale (H&Y).** A staging system used to measure the progression of Parkinson's disease based on the severity of motor symptoms and their impact on daily activities. It ranges from stages 1 to 5, with higher stages indicating more severe impairment.

D.2 Retrieval Methodology

We conducted our search using the folloging keywords:

- **mental health and neurological disorders:** mental health, depression, mental health issues, alzheimer, anxiety, parkinson's, bipolar disorder, stress, dementia, somatic symptoms disorder;
- **speech dataset:** datasets, speech, corpus

As relevant scientific venues, we surveyed the following journals and conference proceedings.

- International Speech Communication Association Archive
- IEEE International Conference on Acoustics, Speech, and Signal Processing
- Language Resources and Evaluation Conference
- ACL Workshop on Computational Linguistics and Clinical Psychology
- International Conference on Machine Learning and Applications
- Clinical NLP Workshop
- ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies
- ACM Multimedia
- IEEE Journal of Biomedical and Health Informatics
- IEEE Transactions on Affective Computing
- IEEE Engineering in Medicine and Biology Society
- American Journal of Speech-Language Pathology
- Frontiers in Neuroscience
- General Psychiatry Journal