This article has been accepted for publication in IEEE Transactions on Neural Networks and Learning Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNNLS.2026.3714047

Multimodal Domain Generalization for Depression Detection: An Attention-Based BiLSTM Network with Domain-Adversarial Training

Ali Tabaraei, Federico Simonetta, and Stavros Ntalampiras

***Abstract*—Automatic depression detection with deep learning has shown promise but often suffers from limited generalization due to domain shift arising from inter-speaker variability. To address this critical issue, we present the first patient-independent multimodal depression detection framework[1] that incorporates domain generalization (DG), jointly leveraging both acoustic and textual modalities. The proposed model integrates *bidirectional Long Short-Term Memory (BiLSTM)* with intra- and cross-modal attention mechanisms, accompanied by segment-level fusion for decision-making. Generalization is further enhanced by applying a *gradient reversal* layer inspired by *Domain-Adversarial Training of Neural Networks (DANN)*, which promotes domain-invariant representations by adversarially limiting the model's ability to identify individual speakers, effectively reducing patient-specific bias. Conducting experiments on the *Androids-Corpus* dataset with a 5-fold cross-validation (CV) protocol, various pairings of audio and text feature extractors were evaluated over different segment durations, determining MelSpec and ItalianBERT as the optimal baseline at a 30-second segment duration. The addition of DG to this baseline yields a 2.5% increase in accuracy and 3.3% in F1-score, achieving 93.2% accuracy, 93.2% precision, 96.2% recall, and 94.2% F1-score, surpassing all existing benchmarks. Extensive ablation studies assess the impact of multimodal fusion, deep architectural choices, and DG, highlighting their combined contribution to robust and generalizable depression detection.**

***Index Terms*—Depression detection, domain generalization, adversarial training, multimodal learning, audio-text analysis.**

I. Introduction

DEPRESSION, one of the most prevalent mental health disorders worldwide [1], can progress to severe outcomes such as self-harm or suicide if left untreated [2]. This risk underscores the urgent need for reliable automated screening methods to complement existing preventive interventions [3].

To enable automatic depression detection, recent advances in Artificial Intelligence (AI) have motivated systematic efforts to collect and analyze behavioral and physiological data, including speech, language, facial expressions, and biological signals, as distinctive indicators of depressive states [4]–[6]. Such data are then typically processed via feature extraction and modeling phases to uncover depression-related patterns.

Among these modalities, *speech* and *linguistic* features have proven to be particularly informative biomarkers of depression.

Ali Tabaraei and Stavros Ntalampiras are with the Department of Computer Science, University of Milan, Milan, Italy. Emails: ali.tabaraei@unimi.it and stavros.ntalampiras@unimi.it

Federico Simonetta is with the Computer Science Department, Gran Sasso Science Institute (GSSI), L'Aquila, Italy. Email: federico.simonetta@gssi.it

Manuscript received 6 October 2025; revised 5 March 2026 and 13 June 2026; accepted 12 July 2026. (Corresponding author: Ali Tabaraei.)

[1]Source code and documentation for this work can be accessed publicly at: https://github.com/tabaraei/MultimodalDG-depression-detection

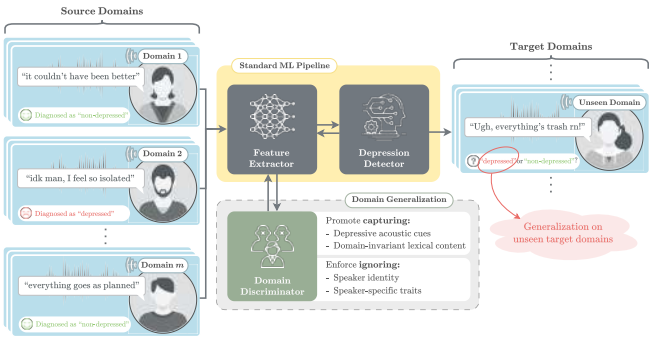

Fig. 1. Overview of the proposed DG framework for multimodal depression detection. Each speaker is represented as a distinct *domain* to model unique variability in audio-textual patterns. The *domain discriminator* constrains the representations learned by the *feature extractor* to suppress speaker-specific traits while retaining depression-related cues. This adversarial setup promotes domain-invariant features and enhances generalization to unseen domains.

Specifically, depressed patients often exhibit recurring acoustic patterns [7], such as reduced pitch variability, slower speaking rate, and longer pauses [8], as well as linguistic traits like self-focused expressions and negative sentiments [9].

Although standard audio-textual machine learning (ML) models have demonstrated improved performance in detecting depression [10], a key barrier to their clinical deployment remains: the lack of subject-independent generalizability. Each individual exhibits unique speech characteristics and language use, which can cause models to capture personal behavioral markers rather than pathological indicators of depression. Consequently, despite strong performance on seen subjects, these models often struggle to generalize to new patients due to inter-speaker variability [11].

To ensure reliability in real-world screening, diagnostic models must be resilient to such variations. Domain adaptation (DA) [12] and domain generalization (DG) [13] are two approaches addressing performance degradation under *domain shift* [14], which occurs when training data (*source domains*) differ from that encountered during inference (*target domains*). While DA requires access to target data during training, DG relies solely on source domains to simulate domain shifts, making it particularly suited to enhancing robustness when diagnosing unseen patients in depression detection [15].

In order to implement the DG paradigm, as shown in Fig. 1, we treat each patient's unique audio-textual data as a separate *domain*, inspired by [16]. Employing this perspective allows us to model the domain shift as inter-speaker variability, with source and target domains corresponding to patients in the training and test sets, respectively.

While a variety of DG methods have been explored in the literature, *domain-invariant representation learning* techniques are extensively studied and proven effective [17]. In particular, these approaches encourage the learning of representations that ignore domain-specific confounding factors while preserving the discriminative patterns of the target task [18].

Within this category, we adopt *domain-adversarial training of neural networks (DANN)* proposed by Ganin *et al.* [19] to guide the standard ML pipeline. Our framework incorporates a *domain discriminator* equipped with a *gradient reversal layer (GRL)* [19], which penalizes the model when it correctly identifies the speaker. This adversarial mechanism drives the feature extractor to learn representations that are invariant to speaker-specific traits, yet remain pathologically informative. It thus facilitates a privacy-preserving, participant-independent depression detector that generalizes to unseen patients.

The proposed framework leverages a *multi-source* DG paradigm [13], jointly modeling acoustic and textual features via a *model-based* fusion strategy [20], [21], all encapsulated within a unified architecture. Our key contributions include:

- Proposing a novel multimodal audio-textual architecture with BiLSTM encoders, intra- and cross-modal attention modules, and segment-level decision-making.
- Conducting a comprehensive analysis of different feature extractor pairs for audio *(MelSpec, HuBERT, Wav2Vec2)* and text *(BERT, ItalianBERT, XLMRoBERTa)* alongside varying segment durations.
- Evaluating the relative contribution of each modality and architectural component with extensive ablation studies.
- Achieving state-of-the-art results of 93.2% accuracy and 94.2% F1-score after incorporating DG, outperforming prior benchmarks despite using 20% less training data.

The rest of this paper is organized as follows: Section II reviews prior studies for comparison, summarizes multimodal approaches, and highlights the relevance of DG. Section III formalizes the problem, providing an introductory context for Section IV, which details different modules of the proposed methodology. Section V presents the experimental setup with a standardized protocol, followed by Section VI, which analyzes the results and presents a thorough ablation study assessing the proposed approach from diverse perspectives. Finally, Sections VIII and VII summarize our conclusions and outline directions for future work.

II. Related Work

This section summarizes prior depression detection studies as a reference for comparison, provides a brief but informative overview of multimodal depression detection, and discusses the relevance of domain generalization in this context.

A. Depression Detection on Androids-Corpus

For comparative purposes in Section VI-C, here we present prior depression detection works on the Androids-Corpus [22] dataset (see Section V-A). This dataset contains recordings of *interview* (I) and *reading* (R) tasks, prompting researchers to adopt a variety of strategies, using these data types either independently or in combination, as outlined in Table I.

TABLE I
OVERVIEW OF DEPRESSION CLASSIFICATION STUDIES ON THE ANDROIDS-CORPUS. SOME STUDIES USED *reading* (R) OR *interview* (I) DATA EXCLUSIVELY, WHILE OTHERS REPORTED RESULTS FOR BOTH IN FUSION ($R \wedge I$) OR SEPARATELY ($R \vee I$), ADOPTING SPEECH-ONLY (S) OR MULTIMODAL (M) APPROACHES. HIGHLIGHTED ROWS ARE NOT COMPARABLE TO OUR STUDY AS WE SOLELY USE THE INTERVIEW DATA.

Features (Model)	Data	Modality
OpenSMILE (SVM, LSTM) [22]	$R \vee I$	S
x-vector + emoHuBERT + TRILLsson (PTM) [23]	$R \vee I$	S
MFCC + TEO + periodicity (HMM + MAP) [24]	$R \vee I$	S
OpenSMILE (LSTM-CDMA) [25]	$R \wedge I$	S
AlexNet (Mixture of Experts) [26]	$R \wedge I$	S
MFCC (Adaptive Knowledge Fusion) [27]	$R \wedge I$	S
STFT spectrograms + ResNet-18 (TFCA) [28]	$R \wedge I$	S
OpenSMILE (CNN-LSTM) / TF-IDF (BERT) [29]	$R \wedge I$	M
OpenSMILE correlations (SVM, LSTM) [30]	R	S
eGeMAPS, TRILLsson4 (XGBoost) [31]	R	S
OpenSMILE (CNN-BiLSTM + RL) [32]	R	S
PDEM embeddings (KELM) [33]	I	S
HuBERT, Wav2Vec2, Whisper (LR) [34]	I	S
DepAudioNet, ResNet, ECAPA-TDNN (MDFA) [35]	I	S
DepAudioNet, ECAPA-TDNN (MTFS-Block) [36]	I	S
Wav2Vec2 (CNN-GRU) / BERT (GRU) [37]	I	M
AlexNet / ItalianBERT (Cross-Attention) [38]	I	M
OpenSMILE, Wav2Vec2 / LIWC (FFN + MV) [39]	I	M
Wav2Vec2 / XLM-RoBERTa (D-CoPE + QF) [40]	I	M

Some studies reported depression detection results on each data type separately ($R \vee I$). The initial baselines by Tao *et al.* [22] utilized OpenSMILE features, with *BS1* employing an SVM and *BS2* an LSTM with segment-level majority voting. Phukan *et al.* [23] proposed *FuSeR*, a deep fusion model with x-vector, TRILLsson, and emoHuBERT embeddings, whereas Ntalampiras [24] achieved competitive performance using an interpretable HMM-based approach with MFCCs, TEO autocorrelation envelopes, and periodicity features.

Conversely, some studies integrated both data types in their experiments ($R \wedge I$). Tao *et al.* [25] developed *Cross-Data Multilevel Attention (CDMA)*, employing attention-enabled LSTMs over OpenSMILE features. Ilias *et al.* [26] stacked log-Mel spectrograms with their first- and second-order derivatives into a 3D representation, subsequently extracting features via AlexNet and processing them with a *Mixture of Experts (MoE)* architecture. Daly and Olukoya [29] processed OpenSMILE features using a hybrid CNN-LSTM while encoding TF-IDF transcripts with BERT, fusing modality-specific outputs at the decision level. Zhou *et al.* [27] utilized an *Adaptive Knowledge Fusion (AKF)* framework that jointly models temporal and channel-domain information from MFCC features. A similar study by Rezaee [28] combined a ResNet-18 backbone and a *Temporal-Frequency-Channel Attention (TFCA)* mechanism to process STFT spectrogram segments.

Other investigations focused solely on a specific data type. Using the reading data (R), Tao *et al.* [30] predicted depression with OpenSMILE-derived correlation representations, whereas Polle *et al.* [31] fed eGeMAPS and TRILLsson4 features into XGBoost to identify confounding biases. With a brain-inspired reinforcement learning architecture, Li *et al.* [32] processed OpenSMILE features via 1-D CNN and BiLSTM modules.

Using the interview data (I), Yu *et al.* [33] fine-tuned a Wav2Vec2-based *PDEM* model and used a *Kernel Extreme Learning Machine (KELM)* for classification, while de Gennes *et al.* [34] benchmarked HuBERT, Wav2Vec2, and Whisper embeddings with logistic regression. Alsenani *et al.* [37] processed Wav2Vec2 audio and BERT text embeddings within a GRU network. Similar to [26], Ilias *et al.* [38] extracted 3D AlexNet-based log-Mel spectrograms and ItalianBERT text features, applying cross-attention instead. Among recent works, Alsarrani *et al.* [39] employed majority voting over audio (OpenSMILE or Wav2Vec2) and LIWC text features. Zhang and Poellabauer [40] proposed Dialogue-based CoPE (D-CoPE) on Wav2Vec2 and XLM-RoBERTa, leveraging an adversarial setup with GRL and dialogue-level transformers. Similarly, Wang *et al.* [35] proposed *Multi-Domain Feature Alignment (MDFA)*, using GRL to learn domain-invariant representations from filter banks via different encoders. In their latest work [36], they used a *Multiple Temporal–Frequency Scale Block (MTFS-Block)* and DWT to enhance DepAudioNet and ECAPA-TDNN baselines.

Most of these studies relied solely on acoustic features, overlooking potential gains from textual information. A further limitation is the lack of a dedicated *validation* set in their experimental design. While [33] used nested cross-validation and [24], [27], [28] reserved a portion of the training data for validation, others trained for a fixed number of epochs, increasing the risk of overfitting and model selection bias.

B. Multimodal Depression Detection

Depression manifests through diverse behavioral signals, making multimodal approaches that integrate audio, text, and visual data well-suited for automated detection. By capturing complementary cues, such methods can typically outperform unimodal systems. Early research by Yang *et al.* [41] fused audio, video, and text descriptors using a deep CNN, while Haque *et al.* [42] later applied a causal CNN to model acoustic, linguistic, and facial features. In a more recent study, Li *et al.* [21] introduced *IISFD*, achieving strong results by integrating visual feature extractors with acoustic and textual features through contrastive learning.

Notably, a major line of research has focused on bimodal audio-text fusion. Early work by Tuka *et al.* [43] modeled audio and text with separate LSTMs, merging the resulting representations in a feedforward network. Later, Makiuchi *et al.* [44] estimated depression status by fusing BERT-based text embeddings with speech features extracted via a pretrained VGG-16 network, further processed using CNN, gated CNN, and LSTM layers. Toto *et al.* [45] proposed *AudiBERT*, a dual self-attentive BiLSTM framework integrating BERT with pretrained audio encoders such as VGGish, SincNet, or Wav2Vec. Shen *et al.* [46] used a GRU-based network on audio Mel spectrograms and a BiLSTM on sentence embeddings, combined through modal attention to predict depression. More recently, Jia *et al.* [47] introduced the *bidirectional multimodal block-recurrent transformer (BMBRT)*, combining BERT and HuBERT for text and audio, respectively. Ding *et al.* [48] then presented *IntervoxNet*, integrating a hybrid BERT-CNN text encoder with *Audio Mel-Spectrogram Transformer (AMST)*. Employing BERT for text and Mel spectrograms for audio, Chen *et al.* [49] achieved promising results with a BiLSTM, cross-attention, and transformer-based fusion network.

The demonstrated success of audio-textual approaches [47], [48] and the effectiveness of intra- and cross-modal attention modules [21], [38] motivated us to adopt a similar strategy, yielding an end-to-end framework that serves as a comparative baseline prior to integrating DG.

C. Domain Generalization

In general, DG has been extensively studied in computer vision, medical imaging, and natural language processing, but its application to speech or multimodal tasks remains limited [13], [17]. In the context of multimodal DG, Zhang *et al.* [50] introduced the *DeVADG* framework, which addresses video-audio domain generalization from a causal perspective by disentangling confounding factors. Planamente *et al.* [51] proposed a novel audio-visual loss function to balance the contributions of both modalities across domains by aligning their feature norms. Dong *et al.* [52] developed *SimMMDG*, a contrastive learning approach on the modality-shared features with distance constraints on modality-specific representations to encourage diversity. In another study, Dong *et al.* [53] presented *MOOSA*, which tackles multimodal open-set domain generalization through self-supervised learning.

Despite these advances, multimodal speech-based DG for depression detection remains largely unexplored. Similar to our approach, Wang *et al.* [35] and Liu *et al.* [54] employed GRL to learn domain-invariant representations, but the former focused on unimodal DA, and the latter conducted unimodal DG using EEG rather than speech. Zhang and Poellabauer [40] is the closest work to ours, applying GRL for multimodal DG. However, they aim to mitigate bias arising from the inclusion of the interviewer's transcripts, while our adversarial setup targets depression-specific features robust to the inter-speaker variability. Unlike previous works, we investigate multimodal DG specifically for patient-independent depression detection, bridging a critical gap in the literature.

III. Problem Formulation

Consider a dataset with an arbitrary number N_i of variable-length audio recordings collected from each speaker i during an interview. These recordings are then preprocessed into K_i fixed-size audio segments, and the corresponding transcripts are obtained. Let $\mathcal{X}$ denote the joint audio-textual feature space and $\mathcal{Y}$ the binary label space for depression. We treat each participant as a distinct *domain*, for whom a joint distribution P_{XY} over $\mathcal{X}\times\mathcal{Y}$ is observed. Given m *source* domains defined as $\{\mathcal{D}_i^{(\mathcal{S})}\}_{i=1}^{m}$, each domain can be represented as:

$$\mathcal{D}_i^{(\mathcal{S})} = \left\{(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)}, d_i, y_i)\right\}_{k=1}^{K_i} \tag{1}$$

where $\mathbf{x}_{i,k}^{(a)}$ and $\mathbf{x}_{i,k}^{(t)}$ represent the *acoustic* and *textual* features extracted by a *modality-specific feature extractor* for segment k of the i-th domain; $d_i \in \{1,\ldots,m\}$ is the source domain label; and $y_i \in \{0,1\}$ indicates the depression status.

During training, a deep neural network referred to as the *multimodal feature extractor (f_θ)* maps each audio-textual pair to a latent representation $\mathbf{z}_{i,k} = f_\theta(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)}) \in \mathbb{R}^d$. This representation is expected to preserve depression-related cues while suppressing speaker-specific characteristics. To this end, the following modules are introduced:

- A *depression detector* (h_ϕ), predicting depression logits $\hat{y}_{i,k} = h_\phi(\mathbf{z}_{i,k}) \in \mathbb{R}$. The detection error is measured via binary cross-entropy, such that computing $\mathrm{BCE}(\hat{y}_{i,k}, y_i)$ across all K_i segments of m source domains yields the overall depression loss $\mathcal{L}_{\mathrm{dep}}(\theta, \phi)$. Consequently, this loss is *minimized* with respect to both θ and ϕ, enabling more accurate depression detection.
- A *domain discriminator* (g_ψ), predicting domain logits $\hat{d}_{i,k} = g_\psi(\mathbf{z}_{i,k}) \in \mathbb{R}^m$. The overall discrimination error $\mathcal{L}_{\mathrm{dom}}(\theta, \psi)$ is defined by computing the cross-entropy loss $\mathrm{CE}(\hat{d}_{i,k}, d_i)$ across all K_i segments of m source domains. While g_ψ aims to *minimize* $\mathcal{L}_{\mathrm{dom}}$ to better identify domains, f_θ *maximizes* it adversarially, tricking g_ψ so that it fails to capture domain-specific features.

This adversarial training strategy promotes domain-invariant representations, thereby enabling generalization to n unseen target domains $\{\mathcal{D}_j^{(\mathcal{T})}\}_{j=1}^n$.

IV. The Proposed Methodology

This section presents an overview of the entire pipeline, elaborating on the key components introduced in Section III and their interplay, specifically the a) preprocessing module, b) audio-text feature extractor specifications, c) multimodal feature extractor architecture, d) domain-adversarial training, and e) inference procedure.

A. Preprocessing and Transcript Extraction

As shown in Fig. 2, each domain comprises N_i disjoint variable-length audio segments, which are first concatenated in chronological order to reconstruct a continuous waveform representing the participant's full speech during the interview. This waveform is then split into K_i segments of a chosen duration, standardizing the input to a fixed number of samples *(segment duration $\times$ sample rate)*. To address the variability in total interview length across domains, if the last split is shorter than 25% of the target duration, it is discarded; otherwise, it is zero-padded to match the intended segment duration.

Each resulting fixed-length audio segment is transcribed using the *"whisper-large-v3"* model, yielding a corresponding textual representation. This process ensures alignment between the acoustic signal and its transcript at the segment level, facilitating effective multimodal analysis.

B. Modality-Specific Feature Extractors

To identify the optimal audio-textual feature extractor pair, different modality-specific feature extractors were evaluated: {*MelSpec, HuBERT, Wav2Vec2*} for audio, alongside {*BERT, ItalianBERT, XLMRoBERTa*} for text. Aside from MelSpec, all transformer models were implemented via Hugging Face, leveraging their last hidden state as frame-level embeddings for each segment. These strategies are detailed in the following.

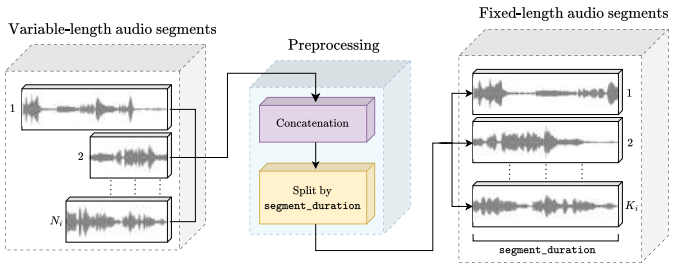

Fig. 2. Preprocessing strategy: Converting N_i variable-length audio segments of the i-th domain into K_i fixed-length segments of a predefined duration.

1) Audio Feature Extractor: Regardless of the method, the extracted acoustic features are standardized into a tensor $\mathbf{x}_{i,k}^{(a)} \in \mathbb{R}^{L^{(a)} \times H^{(a)}}$ for each segment k of the i-th participant, where $L^{(a)}$ denotes the number of frames per segment, and $H^{(a)}$ represents the frame-level feature embedding dimension. In particular, the following methods are evaluated:

- *MelSpec:* Log-Mel spectrograms with FFT size 4096, hop length 512, and 128 filter banks, with first- and second-order derivatives concatenated along the feature axis ($H^{(a)} = 128 \times 3 = 384$).
- *HuBERT: "facebook/hubert-large-ls960-ft"* as the large variant of HuBERT [55] ($H^{(a)} = 1024$).
- *Wav2Vec2: "facebook/wav2vec2-base-960h"* as the base variant of Wav2Vec2 [56] ($H^{(a)} = 768$).

2) Text Feature Extractor: The extracted textual features are represented as $\mathbf{x}_{i,k}^{(t)} \in \mathbb{R}^{L_i^{(t)} \times H^{(t)}}$, where $H^{(t)}$ denotes the token embedding dimension, but $L_i^{(t)}$ is the participant-specific maximum sequence length, to which shorter sequences are zero-padded, since the number of spoken tokens varies across segments of participant i. These embeddings are generated using one of the following pretrained models:

- *BERT: "google-bert/bert-base-multilingual-cased"* as the multilingual variant of BERT [57] ($H^{(t)} = 768$).
- *ItalianBERT: "dbmdz/bert-base-italian-xxl-cased"* as the Italian-specific variant of BERT [58] ($H^{(t)} = 768$).
- *XLMRoBERTa: "FacebookAI/xlm-roberta-large"* as the multilingual variant of RoBERTa [59] ($H^{(t)} = 1024$).

In short, the selected audio-text feature extractor pair converts the preprocessed inputs into $(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})$ for subsequent steps.

C. Multimodal Feature Extractor (f_θ)

This module maps each input pair $(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})$ to a latent feature $\mathbf{z}_{i,k} \in \mathbb{R}^d$ through a structured pipeline of specialized sub-modules. As illustrated in Fig. 3, an independent BiLSTM is first applied to each modality to capture both forward and backward temporal dependencies. This bidirectional approach ensures that the evolving context within both the frame-level acoustic sequences and the token-level textual features is fully captured. Following layer normalization to stabilize training, intra-modal attention highlights distinctive temporal segments within each modality, whereas cross-modal attention captures complementary audio-textual interactions. Together, the core components of f_θ yield the latent representations $\mathbf{z}_{i,k}$, with their architectural configurations detailed in the following.

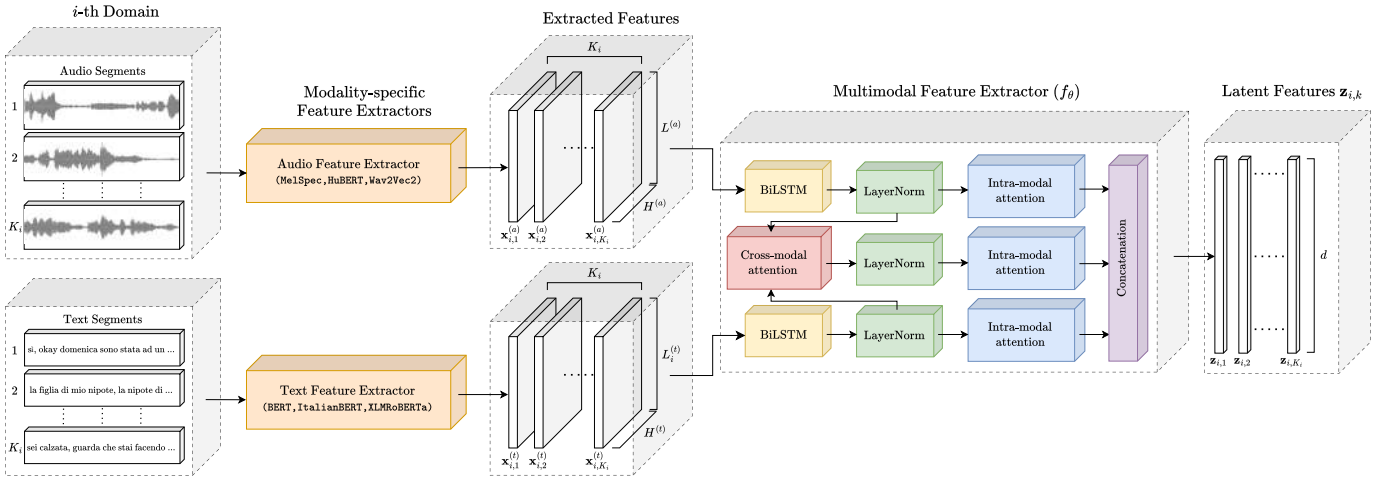

Fig. 3. Feature extraction strategy: For each domain i, the K_i fixed-length audio segments and corresponding transcripts are first fed into their *modality-specific feature extractors* to obtain primary features. Our *multimodal feature extractor* further processes these features, capturing temporal dependencies via BiLSTM and encoding interactions via attention modules. The resulting latent features are then used in an adversarial setup to learn domain-invariant representations.

1) BiLSTM module: Each includes a single hidden layer of size 256, concatenating forward and backward sequences and passing them through a normalization layer. Given $(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})$, the output sequences are:

$$\begin{aligned}\mathbf{seq}_{i,k}^{(a)} &= \mathrm{LayerNorm}\big(\mathrm{BiLSTM}^{(a)}(\mathbf{x}_{i,k}^{(a)})\big) \in \mathbb{R}^{L^{(a)}\times 512}\\ \mathbf{seq}_{i,k}^{(t)} &= \mathrm{LayerNorm}\big(\mathrm{BiLSTM}^{(t)}(\mathbf{x}_{i,k}^{(t)})\big) \in \mathbb{R}^{L_i^{(t)}\times 512}\end{aligned} \tag{2}$$

2) Cross-modal attention (CMA): Let $\alpha = \mathrm{MHA}(Q, K, V)$ be the output of a multi-head attention layer with 4 heads and hidden size 256 that captures interactions among modalities, with the textual sequence as query attending to the aligned acoustic context ($Q = \mathbf{seq}_{i,k}^{(t)}$, and $K = V = \mathbf{seq}_{i,k}^{(a)}$). This output is then passed through a dropout layer ($p = 0.1$), added to the query, and normalized. Formally:

$$\mathbf{seq}_{i,k}^{(c)} = \mathrm{LayerNorm}\big(Q + \mathrm{Dropout}(\alpha)\big) \in \mathbb{R}^{L_i^{(t)}\times 256} \tag{3}$$

3) Intra-modal attention (IMA): A linear attention network with a single hidden layer of size 128 and *Tanh* activation, followed by *Softmax* and a dropout layer ($p = 0.1$) will first compute the attention weights for a given sequence $s \in \mathbb{R}^{L\times H}$. These weights are then applied to the original sequence s via a dot product, yielding a compact representation $v \in \mathbb{R}^H$ that captures its most salient information. We apply this module to the output sequences from Eqs. (2) and (3) as:

$$\begin{aligned}\mathrm{IMA}(s) &= \mathrm{Dropout}\big(\mathrm{Softmax}(\mathrm{Linear}(s))\big)\cdot s\\ \mathbf{v}_{i,k}^{(\alpha)} &= \mathrm{IMA}\big(\mathbf{seq}_{i,k}^{(\alpha)}\big) \in \mathbb{R}^H, \quad \alpha \in \{a, t, c\}\end{aligned} \tag{4}$$

4) Joint feature representation: The final representation is constructed by *concatenating* the attended features in Eq. (4), with $d = 512+512+256 = 1280$ representing the total hidden dimensionality contributed by each component:

$$\mathbf{z}_{i,k} = f_\theta\big(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)}\big) = \Big[\mathbf{v}_{i,k}^{(a)}; \mathbf{v}_{i,k}^{(t)}; \mathbf{v}_{i,k}^{(c)}\Big] \in \mathbb{R}^d \tag{5}$$

D. Domain-Adversarial Training

After extracting the latent features $\mathbf{z}_{i,k}$ using the multimodal feature extractor (f_θ) as defined in Eq. (5), the objective is to ensure that they preserve depression-related information while remaining invariant to inter-speaker variability. To this end, an adversarial framework between a *depression detector* (h_ϕ) and a *domain discriminator* (g_ψ) is introduced, which is described in detail in the following.

1) Depression detector (h_ϕ): The depression detector is implemented as a fully connected network with a single hidden layer of 128 neurons, followed by a *ReLU* activation function and a single output neuron that estimates a depression logit for each segment k, defined as $\hat{y}_{i,k} = h_\phi(\mathbf{z}_{i,k}) \in \mathbb{R}$. The discrepancy between this predicted logit and the ground-truth label for domain i is measured using the binary cross-entropy (BCE) loss during training, computed as:

$$\mathrm{BCE}(\hat{y}_{i,k}, y_i) = -\big[y_i \log\sigma(\hat{y}_{i,k}) + (1-y_i)\log(1-\sigma(\hat{y}_{i,k}))\big] \tag{6}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the sigmoid activation function. Let $\mathcal{L}_{\mathrm{dep}}$ be the overall depression detection loss, defined as the average loss across all segments k and source domains m:

$$\mathcal{L}_{\mathrm{dep}}(\theta, \phi) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\mathrm{BCE}(\hat{y}_{i,k}, y_i)\right) \tag{7}$$

In this setting, the optimization objective is to learn the model parameters θ and ϕ that *minimize* $\mathcal{L}_{\mathrm{dep}}$, enabling accurate prediction of depression for individuals:

$$(\hat{\theta}, \hat{\phi}) \leftarrow \underset{\theta,\phi}{\mathrm{argmin}}\, \mathcal{L}_{\mathrm{dep}}(\theta, \phi) \tag{8}$$

2) Domain discriminator (g_ψ): This module is designed as another fully connected network with a single hidden layer of size 128, followed by a *ReLU* activation function and an output layer with m neurons. Taking as input the same latent features $\mathbf{z}_{i,k}$ for each domain i, it produces $\hat{d}_{i,k} = g_\psi(\mathbf{z}_{i,k}) \in \mathbb{R}^m$ as an output vector over the m possible source domains, such that $\hat{d}_{i,k}[j]$ denotes the logit corresponding to the j-th domain. During training, the prediction error of g_ψ is quantified using the cross-entropy (CE) loss as:

$$\mathrm{CE}(\hat{d}_{i,k}, d_i) = -\log\left(\frac{\exp(\hat{d}_{i,k}[d_i])}{\sum_{j=1}^{m}\exp(\hat{d}_{i,k}[j])}\right) \tag{9}$$

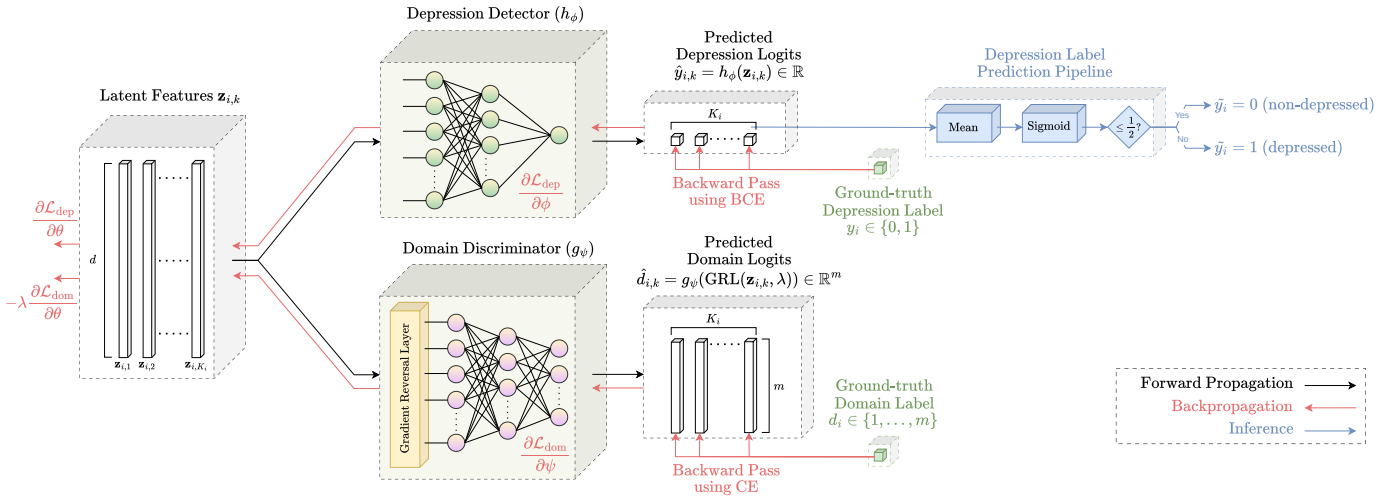

Fig. 4. Adversarial framework between the depression detector (h_ϕ) and the domain discriminator (g_ψ). A gradient reversal layer (GRL) inverts gradients from g_ψ during backpropagation, enabling joint optimization of θ, ϕ, and ψ within a single minimization objective.

where $\exp(z) = e^z$. The overall domain discrimination loss $\mathcal{L}_{\text{dom}}$ is therefore obtained by averaging over all segments k and source domains m:

$$\mathcal{L}_{\text{dom}}(\theta, \psi) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{K_i} \sum_{k=1}^{K_i} \text{CE}(\hat{d}_{i,k}, d_i) \right) \tag{10}$$

Recall that to enable DG, g_ψ is expected to perform poorly at domain discrimination, indicating that f_θ has successfully removed domain-specific information from $\mathbf{z}_{i,k}$. In particular, g_ψ *minimizes* $\mathcal{L}_{\text{dom}}$ to better distinguish source domains, whereas f_θ seeks to *maximize* the same loss, preventing g_ψ from achieving its goal. This adversarial update encourages domain-invariant features to emerge, yielding the following minimax optimization objective:

$$(\hat{\theta}, \hat{\psi}) \leftarrow \underset{\theta}{\text{argmax}} \, \underset{\psi}{\text{argmin}} \, \mathcal{L}_{\text{dom}}(\theta, \psi) \tag{11}$$

To solve this non-trivial optimization via backpropagation, a GRL [19] is integrated into g_ψ, as shown in Fig. 4. The domain logits are obtained as $\hat{d}_{i,k} = g_\psi(\text{GRL}(\mathbf{z}_{i,k}, \lambda)) \in \mathbb{R}^m$, where $\lambda > 0$ is a constant controlling the strength of domain confusion. During the forward pass, GRL acts as the identity function, while in backpropagation it multiplies gradients by $-\lambda$, effectively reversing their direction to $-\lambda \frac{\partial \mathcal{L}_{\text{dom}}}{\partial \theta}$. This transformation enables reformulating Eqs. (8) and (11) into a joint minimization objective:

$$\min_{\theta, \phi, \psi} \; \mathcal{L}_{\text{dep}}(\theta, \phi) + \mathcal{L}_{\text{dom}}(\theta, \psi) \tag{12}$$

The full training pseudocode is presented in Algorithm 1, detailing the joint optimization process that ultimately yields the parameters θ and ϕ used for inference.

E. Inference

To predict the final depression label for an individual *target* domain, the decision relies solely on the domain-invariant representations learned by f_θ and h_ϕ. Consequently, g_ψ can be disabled at inference, as its role is limited to guiding feature learning during adversarial training. Given a target domain i, all its K_i depression logits $\hat{y}_{i,k}$ obtained by the depression detector are averaged to obtain a single score:

$$\bar{y}_i = \frac{1}{K_i} \sum_{k=1}^{K_i} \hat{y}_{i,k} = \frac{1}{K_i} \sum_{k=1}^{K_i} h_\phi\big(\overbrace{f_\theta(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})}^{\mathbf{z}_{i,k}}\big) \in \mathbb{R} \tag{13}$$

and the final predicted depression label is then obtained as:

$$\tilde{y}_i = \begin{cases} 0 \text{ (non-depressed)} & \text{if } \sigma(\bar{y}_i) \leq \frac{1}{2} \\ 1 \text{ (depressed)} & \text{otherwise} \end{cases} \tag{14}$$

V. Experimental Setup

In this section, we describe the experimental framework used to evaluate the proposed multimodal depression detection approach. Specifically, we detail: a) the dataset, including its structure, distribution, and relevance; b) the standardized cross-validation protocol ensuring reproducibility; and c) the training setup, with selected hyperparameter configurations used in the experiments.

A. Dataset

Androids-Corpus [22] is a relatively recent and public Italian dataset, especially noteworthy given the scarcity of depression detection resources in the Italian language. It comprises 228 audio recordings sampled at 16 kHz contributed by 118 native Italian speakers, divided into two groups: 112 recordings of *reading* speech reciting the same text with a total duration of roughly 1 hour and 34 minutes, and 116 recordings of *interview* speech answering a fixed set of open-ended questions spanning around a total of 7 hours and 24 minutes.

Motivated by its spontaneous nature, our study adopts only the interview data, which provides a more realistic setting for speech-based depression detection. Among the corresponding 116 participants, 64 were clinically diagnosed with *depression*, while 52 served as *non-depressed* or *control* participants, with both groups demographically balanced in age and education.

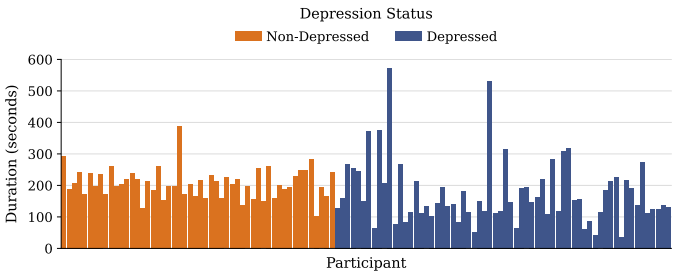

Fig. 5. Waveform length distribution of participants in the Androids-Corpus after concatenating their respective *interview* recordings.

TABLE II
OVERVIEW OF THE HYPERPARAMETER GRID EXPLORED IN THE EXPERIMENTAL SETUP. THE CONFIGURATION ADOPTED FOR THE REPORTED EXPERIMENTS IS HIGHLIGHTED IN BOLD.

Hyperparameter	Candidate Values
MelSpec FFT window size	{2048, **4096**}
MelSpec hop length	{256, **512**, 1024}
MelSpec number of Mel filter banks	{**128**, 192}
Audio BiLSTM hidden dimension	{128, **256**, 512}
Text BiLSTM hidden dimension	{128, **256**, 512}
Intra-modal attention hidden size	{**128**, 256}
Cross-modal attention hidden size	{128, **256**, 512}
Cross-modal attention number of heads	{**4**}
Fully-connected layer hidden size	{64, **128**, 192}
Dropout rate	{**0.1**, 0.2}
GRL scaling coefficient (λ)	{0.3, **0.5**, 0.7, 1}
Learning rate (η)	{**1**, 2, 5}$\times$**10**$^{-5}$
AdamW optimizer weight decay	{**1**, 2, 5}$\times$**10**$^{-5}$
ReduceLROnPlateau scheduler decay factor (γ)	{0.1, 0.2, **0.5**}
ReduceLROnPlateau scheduler patience epochs (P_{schd})	{**1**, 2, 3, 5}
Early stopping epsilon (ε)	{**0.1**, 0.2}
Early stopping patience epochs (P_{stop})	{2, **3**, 5}
Maximum training epochs (N)	{**100**}

The dataset also provides a segmentation of the audio files to ease the process of eliminating the interviewer's speech. As shown in Fig. 5, the duration of the concatenated recordings for each individual ranges approximately from 40 seconds to just under 600 seconds.

B. k-fold Cross-Validation

Replicating the data splits introduced by the authors of the dataset, the experiments in this study adopted a k-fold ($k = 5$) cross-validation protocol. These splits keep all the participant's recordings in the same fold to ensure no overlap between train and test sets, which prevents memorizing speaker-specific traits. At each iteration $k = 1, \ldots, 5$, the k-th fold was treated as the test set, while the remaining data were further divided using a stratified split to maintain the distribution of labels across splits, with 80% used for training and 20% reserved for validation. As the dataset was almost balanced in terms of depressed vs non-depressed classes, standard evaluation metrics including accuracy, precision, recall, and F1-score were used for performance evaluation.

To ensure robustness and account for potential variability, the experimental results reported in Section VI-A are presented as the mean and standard deviation over *three* independent runs of each experiment. In contrast, the results in Sections VI-B and VI-D are reported as the mean and standard deviation across folds from a single run, in line with prior studies using the Androids-Corpus to ensure fair comparison and reliable ablation analysis.

Algorithm 1 MultimodalDG Training

Require: Learning rate η, GRL coefficient λ, max epochs N
Require: m source domains $\{\mathcal{D}_i^{(\mathcal{S})}\}_{i=1}^{m}$, each comprising K_i fixed-size segments $\{(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)}, d_i, y_i)\}_{k=1}^{K_i}$

Split $\{\mathcal{D}_i^{(\mathcal{S})}\}_{i=1}^{m}$ into training set S and validation set V
Initialize randomly the model parameters θ, ϕ, ψ
for epoch $= 1$ **to** N **do**
  $\mathcal{L}_{\text{train}} \leftarrow 0$, $\mathcal{L}_{\text{val}} \leftarrow 0$

  # Training Phase
  for each domain i **in** S **do**
    for segment $k = 1$ **to** K_i **do**
      Extract the latent features $\mathbf{z}_{i,k} \leftarrow f_\theta(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})$
      Predict depression logits $\hat{y}_{i,k} \leftarrow h_\phi(\mathbf{z}_{i,k})$
      Predict domain logits $\hat{d}_{i,k} \leftarrow g_\psi(\text{GRL}(\mathbf{z}_{i,k}, \lambda))$
      $\ell_{i,k} \leftarrow \text{BCE}(\hat{y}_{i,k}, y_i) + \text{CE}(\hat{d}_{i,k}, d_i)$
    end for
    Average the loss of segments $\mathcal{L}_i \leftarrow \frac{1}{K_i} \sum_{k=1}^{K_i} \ell_{i,k}$
    Update θ, ϕ, ψ to minimize $\mathcal{L}_i$ using $\text{AdamW}(\eta)$
    $\mathcal{L}_{\text{train}} \leftarrow \mathcal{L}_{\text{train}} + \mathcal{L}_i$
  end for

  # Validation Phase
  for each domain i **in** V **do**
    for segment $k = 1$ **to** K_i **do**
      Extract the latent features $\mathbf{z}_{i,k} \leftarrow f_\theta(\mathbf{x}_{i,k}^{(a)}, \mathbf{x}_{i,k}^{(t)})$
      Predict depression logits $\hat{y}_{i,k} \leftarrow h_\phi(\mathbf{z}_{i,k})$
      $\ell_{i,k} \leftarrow \text{BCE}(\hat{y}_{i,k}, y_i)$
    end for
    $\mathcal{L}_{\text{val}} \leftarrow \mathcal{L}_{\text{val}} + \frac{1}{K_i} \sum_{k=1}^{K_i} \ell_{i,k}$
  end for

  if $\text{EarlyStopping}(\mathcal{L}_{\text{train}}, \mathcal{L}_{\text{val}})$ is triggered **then**
    break
  end if
  Update η using $\text{ReduceLROnPlateau}(\mathcal{L}_{\text{val}})$
end for
return θ, ϕ

C. Training Specifications

Model training was performed using the *AdamW* optimizer, with learning rate (η) and weight decay each set to 1×10^{-5}. Using PyTorch's ReduceLROnPlateau scheduler with decay factor $\gamma = 0.5$ and patience $P_{\text{schd}} = 1$ epochs, η was adjusted dynamically in response to plateaus in the validation loss.

As outlined in Algorithm 1, training and validation losses were monitored using a custom $\text{EarlyStopping}(\mathcal{L}_{\text{train}}, \mathcal{L}_{\text{val}})$ mechanism, which terminated training if: (i) changes in $\mathcal{L}_{\text{train}}$ remained below $\varepsilon = 0.1$ for $P_{\text{stop}} = 3$ consecutive epochs, indicating convergence, or (ii) either $\mathcal{L}_{\text{train}}$ or $\mathcal{L}_{\text{val}}$ maintained an increasing trend over the same number of epochs. Although the maximum number of epochs was set a priori to $N = 100$,

TABLE III
PERFORMANCE COMPARISON OF AUDIO-TEXT FEATURE EXTRACTOR PAIRINGS ACROSS DIFFERENT SEGMENT DURATIONS. EACH CELL REPORTS ACCURACY / PRECISION / RECALL / F1-SCORE, PRESENTED IN THE SAME ORDER AS PERCENTAGES (%).

Audio Feature Extractor	Text Feature Extractor	Segment Duration			
		20s	30s	45s	60s
MelSpec	BERT	77.2 / 76.2 / 88.1 / 80.5	85.2 / 85.9 / 89.3 / 86.8	84.3 / 84.9 / 87.3 / 85.4	85.8 / 87.0 / **88.9** / 87.1
	ItalianBERT	86.4 / **86.8** / 89.5 / 87.3	**90.4** / **90.8** / 92.0 / **90.8**	**88.1** / **87.1** / **90.9** / **88.7**	**87.5** / **90.4** / 87.8 / **88.0**
	XLMRoBERTa	68.5 / 66.1 / 92.7 / 76.1	67.9 / 69.1 / 85.4 / 74.2	70.8 / 69.8 / 85.6 / 75.0	75.5 / 75.0 / 86.2 / 78.9
HuBERT	BERT	76.2 / 74.1 / 86.4 / 78.8	77.9 / 76.9 / 88.0 / 80.9	77.3 / 77.5 / 87.4 / 80.4	79.9 / 81.1 / 85.4 / 82.0
	ItalianBERT	83.1 / 80.4 / 91.7 / 85.0	82.5 / 80.6 / 91.1 / 84.7	85.4 / 85.7 / 89.5 / 86.6	80.2 / 80.8 / 87.1 / 82.5
	XLMRoBERTa	70.4 / 71.4 / 81.6 / 72.8	71.3 / 72.4 / 87.7 / 76.0	70.8 / 71.7 / 83.2 / 74.7	76.4 / 75.5 / 86.5 / 78.6
Wav2Vec2	BERT	78.8 / 77.6 / 86.6 / 80.9	81.0 / 79.6 / 90.4 / 83.8	79.7 / 84.5 / 82.8 / 80.4	80.8 / 83.6 / 83.0 / 82.1
	ItalianBERT	**86.8** / 85.4 / **93.5** / **88.3**	86.2 / 85.9 / **92.2** / 88.1	83.1 / 85.5 / 87.7 / 85.1	84.0 / 85.1 / 87.3 / 85.5
	XLMRoBERTa	73.3 / 73.3 / 79.6 / 75.0	73.9 / 75.0 / 82.7 / 76.6	71.0 / 71.5 / 81.8 / 74.3	81.6 / 81.1 / 86.1 / 82.7

training typically halted after no more than 35 epochs in most experiments, effectively preventing overfitting due to the rising validation loss.

Different hyperparameters in this study were chosen via empirical tuning over the search space summarized in Table II, with the final optimal values highlighted in bold for reference.

The proposed approach was developed using Python 3.12.8, built on the *PyTorch* framework, and executed on a server with two *NVIDIA TITAN V* GPUs (each with 12 GB of VRAM, running CUDA version 12.8).

VI. RESULTS

This section presents the experimental results of our study, specifically detailing: a) an analysis of feature extractors across segment durations, b) the impact of domain generalization, c) a comparison with state-of-the-art methods, and d) ablation studies on modalities and architectural variants.

A. Feature Extractors vs. Segment Durations

In order to find a competitive baseline, different selections of modality-specific feature extractor pairs and segment durations for fixed-length audio segments were evaluated in a structured set of experiments. In particular, every possible pairing of three audio extractors (MelSpec, HuBERT, and Wav2Vec2) with three text extractors (BERT, ItalianBERT, and XLMRoBERTa) across segment durations of 20, 30, 45, and 60 seconds were examined for this purpose.

As summarized in Table III, the pairing of MelSpec feature extractor for audio and ItalianBERT for the text modality at a 30-second segment duration achieved the highest performance. With an average accuracy and F1-score of 90.4% and 90.8%, respectively, this combination was selected as the baseline for downstream analysis.

To assess the impact of segment duration, Fig. 6 presents average performance across all feature extractor combinations for each duration, derived from Table III. Overall, adopting longer speech segments improves accuracy, though at the cost of higher GPU memory usage for processing larger sequences. Given this trade-off, a 30-second segment duration provides a good balance between performance and computational cost.

Alternatively, Fig. 7 illustrates the average results of each audio-textual feature extractor. The combination of MelSpec (audio) and ItalianBERT (text) consistently yields the highest performance, likely due to MelSpec's compact representations and the Italian-specific design of ItalianBERT.

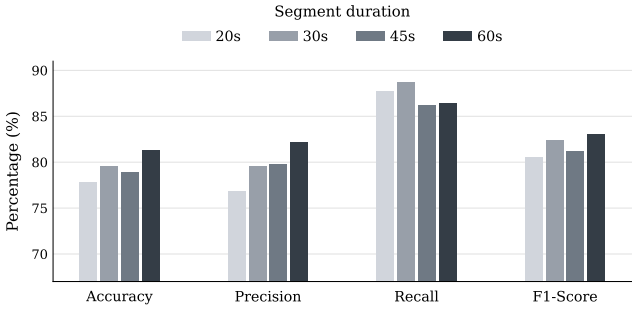

Fig. 6. Average performance across all experiments for each segment duration, computed across all the audio-text pairings for each duration in Table III.

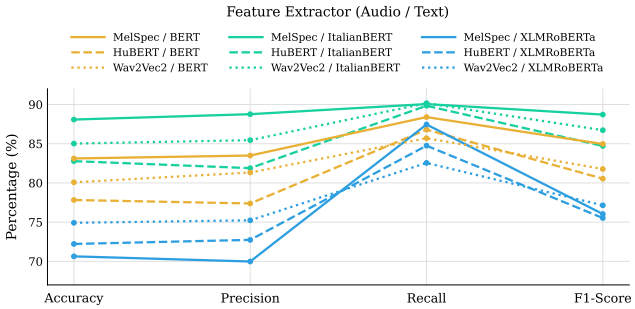

Fig. 7. Average performance of modality-specific feature extractor pairings, computed across all segment durations for each pairing from Table III.

B. Domain Generalization Effect

The previously identified baseline model was extended by integrating the domain discriminator g_ψ to assess the proposed DG framework. This inclusion led to an improvement of 2.5% in accuracy and 3.3% in F1-score, yielding a final performance of 93.2% accuracy, 93.2% precision, 96.2% recall, and 94.2% F1-score, highlighting the contribution of adversarial training in improving generalization to unseen target domains.

Analyzing these performance gains, McNemar's test with continuity correction was applied to individual predictions. Given roughly 23 testing samples per fold in the standardized

TABLE IV
PERFORMANCE COMPARISON WITH PRIOR STUDIES ON THE ANDROIDS-CORPUS FOR DEPRESSION DETECTION, REPORTED AS PERCENTAGES (%).

Study	Model	Features	Accuracy	Precision	Recall	F1-score
[22] Rand.	Random Classifier	MFCC, RMSE, ZCR, etc. (OpenSMILE)	50.5	55.2	55.2	55.2
[22] BS1	SVM	MFCC, RMSE, ZCR, etc. (OpenSMILE)	73.3 ± 10.6	73.5 ± 16.1	74.5 ± 13.2	73.6 ± 13.6
[22] BS2	LSTM	MFCC, RMSE, ZCR, etc. (OpenSMILE)	83.9 ± 1.3	85.8 ± 3.1	86.1 ± 2.7	84.7 ± 0.9
[23] FuSeR	PTM	x-vector, emoHuBERT, TRILLsson	87.9	–	–	87.8
[24] UBM-HMM	HMM and MAP	MFCC, TEO, periodicity	87.0 ± 1.2	85.8 ± 1.9	92.3 ± 1.8	88.9 ± 1.7
[33] PDEM	KELM	PDEM embeddings (based on Wav2Vec2)	88.2 ± 8.5	89.6 ± 6.5	84.4 ± 17.9	86.3 ± 11.5
[34] HuBERT L	Logistic Regression	HuBERT-L embeddings	–	–	–	92.0
[35] MDFA	MDFA +GRL	ResNet-18	–	–	–	79.2
[36] MTFS	MTFS-Block	ECAPA-TDNN	–	79.4	78.8	77.4
[37] Multimodal	GRU	Wav2Vec2 / BERT	86.2 ± 7.8	–	–	–
[38] Concatenation	Cross-Attention	AlexNet (3D log-Mel) / ItalianBERT	91.5 ± 6.1	91.5 ± 8.7	93.4 ± 6.0	92.1 ± 5.5
[39] MIL	Majority-voting	OpenSMILE / LIWC	92.5 ± 1.1	–	–	93.1 ± 1.1
[40] Full model	D-CoPE + QF +GRL	Wav2Vec2 / XLM-RoBERTa	–	–	–	73.0
This work	MultimodalDG	MelSpec / ItalianBERT	$\mathbf{93.2 \pm 5.7}$	$\mathbf{93.2 \pm 9.7}$	$\mathbf{96.2 \pm 4.7}$	$\mathbf{94.2 \pm 4.7}$

TABLE V
ABLATION STUDY HIGHLIGHTING THE CONTRIBUTION OF EACH MODALITY TO MODEL PERFORMANCE.

Modality	Accuracy	Precision	Recall	F1-Score
Multimodal	$\mathbf{93.2 \pm 5.7}$	$\mathbf{93.2 \pm 9.7}$	$\mathbf{96.2 \pm 4.7}$	$\mathbf{94.2 \pm 4.7}$
Audio-only	73.4 ± 9.5	74.9 ± 10.1	79.0 ± 10.7	76.2 ± 8.3
Text-only	85.3 ± 8.1	87.7 ± 9.7	87.7 ± 9.4	87.0 ± 6.0

TABLE VI
ABLATION STUDY ANALYZING THE IMPACT OF ALTERNATIVE DESIGN CHOICES ON MODEL PERFORMANCE. ABBREVIATIONS INCLUDE: DOMAIN GENERALIZATION (DG), INTRA-MODAL ATTENTION (IMA), CROSS-MODAL ATTENTION (CMA), LAYER NORMALIZATION (LN).

Architecture	Accuracy	Precision	Recall	F1-Score
Full Model	$\mathbf{93.2 \pm 5.7}$	$\mathbf{93.2 \pm 9.7}$	$\mathbf{96.2 \pm 4.7}$	$\mathbf{94.2 \pm 4.7}$
No DG	90.7 ± 8.9	90.4 ± 10.4	91.9 ± 7.9	90.9 ± 8.5
No IMA	87.3 ± 10.4	85.9 ± 12.9	92.6 ± 6.5	88.7 ± 8.9
No CMA	89.8 ± 6.8	89.3 ± 10.2	92.5 ± 6.9	90.4 ± 6.4
No LN	88.9 ± 7.7	89.5 ± 12.1	92.6 ± 6.9	90.2 ± 6.3

experimental protocol, achieving $p < 0.05$ requires substantial disagreement between models. Hence, the null hypothesis was not rejected, reflecting that insufficient sample size constrains statistical significance, even with strong observed results.

To provide an alternative perspective on model performance, we report the cumulative confusion matrix in Fig. 8, obtained by summing the per-fold confusion matrices across the 5-fold cross-validation. Unlike the averaged metrics reported earlier, this summation conceptually reflects the model's behavior over the entire dataset. The proposed model correctly identifies 95.3% and 90.4% of depressed and non-depressed individuals, respectively. In addition, it achieves a low false negative rate (FNR $= 4.7\%$), which is crucial in clinical screening to avoid missing individuals at risk while maintaining strong specificity for the non-depressed group.

C. Comparison with Androids-Corpus Benchmarks

Table IV benchmarks the proposed approach against prior studies on the Androids-Corpus dataset (see Section II-A for reference). The proposed model outperforms all these efforts across every evaluation metric, attaining a peak F1-score of 94.2%, which represents a clear improvement over the 93.1% of the closest competitor [39]. This comparison validates the strength and reliability of the proposed multimodal framework, and the effectiveness of the domain generalization strategy in capturing depression-specific feature embeddings.

D. Ablation Studies

A set of ablation experiments was carried out to investigate the contribution of each component in the proposed strategy.

First, the role of each modality was evaluated, as reported in Table V. The highest performance was achieved with the full multimodal setup, demonstrating the complementary value of integrating both acoustic and linguistic modalities for accurate depression detection.

The large performance drop in unimodal experiments (e.g., decreasing to 73.4% accuracy for audio-only) is directly tied to the forced architectural adaptations within f_θ. When restricted to a single modality, the network excludes not only the intra-modally attended BiLSTM stream of the omitted modality (see Fig. 3), but also inherently disables the CMA module and its subsequent IMA block, as there is no complementary modality to drive the attention computation. Together, these alterations justify the substantial performance gap when compared to the full multimodal framework.

Next, a series of architectural modifications were applied to assess the contribution of each component, as summarized in Table VI. The results show that disabling any of these modules inevitably degrades performance, thereby confirming the critical role each plays and underscoring the importance of attention mechanisms, normalization layers, and particularly domain generalization in achieving optimal results.

Notably, the contribution of IMA to the overall performance outweighs that of CMA. IMA highlights salient audio-textual segments to assist subsequent layers in filtering uninformative temporal noise from the BiLSTM and CMA outputs, playing a critical role in achieving optimal performance. In contrast, the architecture is less sensitive to the removal of the cross-modal

sequence $\mathbf{seq}_{i,k}^{(c)}$ (see Eqs. (3)–(5)) from the final latent feature $\mathbf{z}_{i,k}$. This suggests that the subsequent feed-forward network h_ϕ can partially compensate for the absence of CMA, likely by relying on the remaining unimodal branches.

VII. Limitations and Future Work

Despite its methodological contributions and strong results, this study leaves several promising directions for future work. First, due to computational constraints, the modality-specific feature extraction was performed as a stand-alone stage prior to model training. Consequently, the framework is limited by the absence of a fine-tuning stage for the transformer-based encoders, which may reduce performance compared to fully end-to-end approaches. Second, the experiments are limited by the use of a single dataset, due to the scarcity of high-quality, publish-ready datasets in Italian. Future work should explore cross-dataset analysis to confirm the framework's robustness and generalizability, as more datasets become available. Lastly, interpreting internal embeddings under DG settings remains an open challenge, motivating the integration of explainable AI to strengthen clinical trust and enhance the robustness of mental health monitoring systems.

VIII. Conclusion

This study is the first effort to adopt domain generalization for learning domain-invariant features robust to inter-speaker variability in multimodal depression detection. It implements a novel architecture that combines BiLSTM sequence modeling, attention mechanisms, and segment-level decision-making. To identify the most effective audio-text feature extractor pairing, a series of experiments were conducted across various segment durations, ultimately finding the combination of MelSpec and ItalianBERT with a 30-second duration to be the most effective configuration. Next, this baseline model was integrated with domain-adversarial training, achieving notable improvements in classification performance. The obtained results surpassed all previous benchmarks on the Androids-Corpus, despite the reduced training set size due to validation splitting. Lastly, comprehensive ablation studies confirmed the importance of integrating both modalities and the critical role of each model component. Overall, these findings underscore the applicability of DG as a powerful tool to boost real-world generalizability, enabling patient-independent diagnosis.

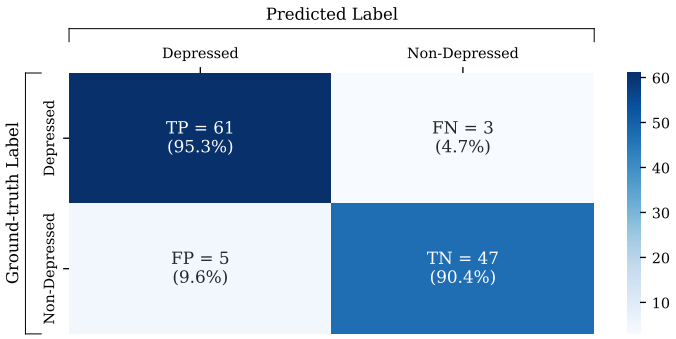

Fig. 8. Cumulative confusion matrix of the proposed model, compiled by summing each fold's confusion matrix across the 5-fold cross-validation setup.

References

[1] S. Shorey, E. D. Ng, and C. H. Wong, "Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis," *British journal of clinical psychology*, vol. 61, no. 2, pp. 287–305, 2022.

[2] V. Baldini, M. Gnazzo, M. Maragno, R. Biagetti, C. Stefanini, F. Canulli, G. Varallo, C. Donati, G. Neri, A. Fiorillo *et al.*, "Suicidal risk among adolescent psychiatric inpatients: the role of insomnia, depression, and social-personal factors," *European Psychiatry*, vol. 68, no. 1, p. e42, 2025.

[3] L. Xia, Y. Feng, Z. Guo, J. Ding, Y. Li, Y. Li, M. Ma, G. Gan, Y. Xu, J. Luo, Z. Shi, and Y. Guan, "MuLHiTA: A Novel Multiclass Classification Framework With Multibranch LSTM and Hierarchical Temporal Attention for Early Detection of Mental Stress," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9657–9670, Dec. 2023.

[4] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by intelligent speech signal processing: A systematic survey," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 701–711, 2023.

[5] L. Qu, C. Weber, W. Wang, J. Jin, Y. Gao, T. Li, and S. Wermter, "Disentanglement of Prosody Representations via Diffusion Models and Scheduled Gradient Reversal," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 8, pp. 15 043–15 054, Aug. 2025.

[6] W. Yuan, X. Zhang, X. Zhang, S. Wang, T. Wang, T. Zhang, Q. Zhao, and B. Hu, "Discovery of shared latent nonlinear effective connectivity for eeg-based depression detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 10 663–10 677, 2025.

[7] S. Koops, S. G. Brederoo, J. N. de Boer, F. G. Nadema, A. E. Voppel, and I. E. Sommer, "Speech as a biomarker for depression," *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, vol. 22, no. 2, pp. 152–160, 2023.

[8] Y. Di, E. Rahmani, J. Mefford, J. Wang, V. Ravi, A. Gorla, A. Alwan, K. S. Kendler, T. Zhu, and J. Flint, "Unraveling the associations between voice pitch and major depressive disorder: a multisite genetic study," *Molecular Psychiatry*, vol. 30, no. 6, pp. 2686–2695, 2025.

[9] L. Gu, M. Li, and Y. Li, "Linguistic markers of depression and emergent self-stigma in online self-disclosures: A mixed-methods study on chinese social media," *Journal of Affective Disorders*, p. 120765, 2025.

[10] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–41, 2023.

[11] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2022.

[12] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[13] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[14] M. Jahanifar, M. Raza, K. Xu, T. T. L. Vuong, R. Jewsbury, A. Shephard, N. Zamanitajeddin, J. T. Kwak, S. E. A. Raza, F. Minhas, and N. Rajpoot, "Domain Generalization in Computational Pathology: Survey and Guidelines," *ACM Comput. Surv.*, vol. 57, no. 11, pp. 285:1–285:37, Jun. 2025.

[15] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards Out-Of-Distribution Generalization: A Survey," *ArXiv*, Aug. 2021.

[16] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *arXiv preprint arXiv:1804.10745*, 2018.

[17] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to Unseen Domains: A Survey on Domain Generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, Aug. 2023.

[18] C. Rohlfs, "Generalization in neural networks: A broad survey," *Neurocomputing*, vol. 611, p. 128701, Jan. 2025.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[20] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024.

[21] M. Li, Y. Wei, Y. Zhu, S. Wei, and B. Wu, "Enhancing multimodal depression detection with intra-and inter-sample contrastive learning," *Information Sciences*, vol. 684, p. 121282, 2024.

[22] F. Tao, A. Esposito, and A. Vinciarelli, "The androids corpus: A new publicly available benchmark for speech based depression detection," *Depression*, vol. 47, pp. 11–9, 2023.

[23] O. C. Phukan, S. R. Behera, S. Singh, M. Singh, V. Rajan, A. B. Buduru, R. Sharma, and S. Prasanna, "Avengers assemble: Amalgamation of non-semantic features for depression detection," *arXiv preprint arXiv:2409.14312*, 2024.

[24] S. Ntalampiras, "Interpretable probabilistic identification of depression in speech," *Sensors*, vol. 25, no. 4, p. 1270, 2025.

[25] F. Tao, X. Ge, W. Ma, A. Esposito, and A. Vinciarelli, "Cross-data multilevel attention for depression detection: Analyzing the interplay between read and spontaneous speech," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 1169–1176.

[26] L. Ilias and D. Askounis, "Mixture of experts for recognizing depression from interview and reading tasks," *arXiv preprint arXiv:2502.20213*, 2025.

[27] L. Zhou, X. Zhang, S. Guan, and X. Luo, "Adaptive Knowledge Fusion Model for Depression Recognition," *IEEE Transactions on Computational Social Systems*, 2025.

[28] K. Rezaee, "Depression detection from speech data using deep learning–based optimized temporal–frequency–channel attention with interpretable acoustic–prosodic mapping," *Journal of Affective Disorders*, p. 121077, 2026.

[29] K. Daly and O. Olukoya, "Depression detection in read and spontaneous speech: A Multimodal approach for lesser-resourced languages," *Biomedical Signal Processing and Control*, vol. 108, p. 107959, 2025.

[30] F. Tao, W. Ma, X. Ge, A. Esposito, and A. Vinciarelli, "The relationship between speech features changes when you get depressed: Feature correlations for improving speed and performance of depression detection," *arXiv preprint arXiv:2307.02892*, 2023.

[31] R. Polle, S. Fara, A. Georgescu, S. Goria, and N. Cummins, "Revealing confounding biases: A novel benchmarking approach for aggregate-level performance metrics in health assessments," *Accepted for Interspeech 2024, Kos Island, Greece*, 2024.

[32] D. Li, J. Yao, Z. Wang, and Y. Yi, "FAD3QN: A Brain-Inspired Deep Reinforcement Learning Model for Speech Depression Detection," *IEEE Transactions on Computational Social Systems*, 2025.

[33] J. Yu and H. Kaya, "Using emotionally rich speech segments for depression prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[34] M. de Gennes, A. Lesage, M. Denais, X.-N. Cao, S. Chang, P. Van Remoortere, C. Dakhlia, and R. Riad, "Probing mental health information in speech foundation models," *arXiv preprint arXiv:2409.19042*, 2024.

[35] M. Wang, S. Kato, W. Gu, J. Yan, and F. Ren, "Cross-Language Depression Detection Based on Multi-Domain Feature Alignment," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2025, pp. 1–7.

[36] M. Wang, S. Kato, W. Gu, F. Ren, and J. Yan, "Depression detection from speech signals using a multiple temporal-frequency scale Channel Attention Mechanism," *Biomedical Signal Processing and Control*, vol. 113, p. 108873, 2026.

[37] B. Alsenani, A. Esposito, A. Vinciarelli, and T. Guha, "Assessing privacy risks of attribute inference attacks against speech-based depression detection system," *ECAI*, 2024.

[38] L. Ilias and D. Askounis, "A cross-attention layer coupled with multimodal fusion methods for recognizing depression from spontaneous speech," in *Proc. Interspeech 2024*, 2024, pp. 912–916.

[39] R. Alsarrani, A. Esposito, and A. Vinciarelli, "Punctual or Continuous? Analyzing Depression Traces in Language and Paralanguage with Multiple Instance Learning," in *Proceedings of the 27th International Conference on Multimodal Interaction*. Canberra Australia: ACM, Oct. 2025, pp. 614–623.

[40] E. Zhang and C. Poellabauer, "Mitigating Interviewer Bias in Multimodal Depression Detection: An Approach with Adversarial Learning and Contextual Positional Encoding," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 12 169–12 188.

[41] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, 2017, pp. 53–59.

[42] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.

[43] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[44] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[45] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 4145–4154.

[46] Y. Shen, H. Yang, and L. 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)*. IEEE, 2022, pp. 6247–6251.

[47] X. Jia, X. Zhao, B. Tang, and R. Jiang, "Bidirectional multimodal block-recurrent transformers for depression detection," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 3323–3328.

[48] H. Ding, Z. Du, Z. Wang, J. Xue, Z. Wei, K. Yang, S. Jin, Z. Zhang, and J. Wang, "Intervoxnet: a novel dual-modal audio-text fusion network for automatic and efficient depression detection from interviews," *Frontiers in Physics*, vol. 12, p. 1430035, 2024.

[49] Z. Chen, D. Wang, L. Lou, S. Zhang, X. Zhao, S. Jiang, J. Yu, and J. Xiao, "Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition," *Information Fusion*, vol. 117, p. 102861, 2025.

[50] S. Zhang, X. Feng, W. Fan, W. Fang, F. Feng, W. Ji, S. Li, L. Wang, S. Zhao, Z. Zhao *et al.*, "Video-audio domain generalization via confounder disentanglement," in *Proceedings of the AAAI conference on artificial intelligence*, ser. 37, no. 12, 2023, pp. 15 322–15 330.

[51] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1807–1818.

[52] H. Dong, I. Nejjar, H. Sun, E. Chatzi, and O. Fink, "Simmmdg: A simple and effective framework for multi-modal domain generalization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 674–78 695, 2023.

[53] H. Dong, E. Chatzi, and O. Fink, "Towards multimodal open-set domain generalization and adaptation through self-supervision," in *European Conference on Computer Vision*. Springer, 2024, pp. 270–287.

[54] S. Liu, L. An, and Z. Jia, "A domain adversarial learning framework for major depression disorder diagnosis," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[55] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[56] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[58] S. Schweter, "Italian bert and electra models," Available at https://doi.org/10.5281/zenodo.4263142, 2020, version 1.0.1, Zenodo.

[59] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

Ali Tabaraei is a Ph.D. candidate in Computer Science at the University of Milan, Italy, where he also received his master's degree in the same field in 2025. His current doctoral research focuses on the generalizability and interpretability aspects of advanced deep neural networks applied to health acoustics. Particularly, his research interests include Domain Generalization, Domain Adaptation, Audio Pattern Recognition, Explainable AI, and Health AI.

Federico Simonetta is a post-doctoral researcher in the Laudare ERC AdG project at the Gran Sasso Science Institute (GSSI). He previously worked as a post-doctoral researcher at the Universidad Complutense de Madrid and Instituto Complutense de Ciencias Musicales (ICCMU) in the Didone ERC AdG project. He obtained his Ph.D. in Computer Science from the University of Milan in 2022. He is active in the scientific committees of several international journals and conferences. His main research interests are music information processing, machine learning, audio processing, and handwritten music/text recognition.

Stavros Ntalampiras is an Associate Professor at the Department of Computer Science, University of Milan, Italy. He received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2010, respectively. He has carried out research and/or didactic activities at Politecnico di Milano, the Joint Research Center of the European Commission, the National Research Council of Italy, and Bocconi University. Currently, he is an Associate Editor of IEEE TNNLS, PLOS One, IET Signal Processing and CAAI Transactions on Intelligence Technology, as well as member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing. His research interests include content-based signal processing, machine learning, audio pattern recognition, bioacoustics, health acoustics, and cyber-physical systems.