university of groningen

campus fryslân

Layer-wise Cross-Lingual Depression Detection from Speech: A HuBERT-Based Study on English and Mandarin

Hang Chen

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Layer-wise Cross-Lingual Depression Detection from Speech: A HuBERT-Based Study on English and Mandarin

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr.Vass Verkhodanova (Voice Technology, University of Groningen)

Hang Chen (S-5944562)

July 23, 2025

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Vass Verkhodanova, for her guidance, insightful feedback, and continuous encouragement throughout the thesis project. Her support was crucial for both the technical and academic development of this work.

I would also like to thank all the professors at the Department of Speech Technology at the University of Groningen, who provided a supportive and intellectually stimulating environment during my studies. I would also like to thank the faculty who laid the foundation for this research.

Special thanks to my classmates, especially those who shared their views during the long process of studying and writing the thesis - your collaboration made this journey both productive and enjoyable.

I thank the Center for Information Technology at the University of Groningen for technical support and for providing access to the Hábrók High Performance Computing Cluster, which was essential for the large-scale processing and model training phases of this research.

Finally, I would like to thank my family and friends for their patience, emotional support, and belief in me. Without their encouragement, this paper would not have been possible.

Abstract

Depression is a global mental health challenge. While many recent studies have applied self-supervised learning (SSL) models for speech-based depression detection, most are trained and evaluated in monolingual settings, predominantly in English [1]. This study investigates whether depression-related acoustic cues are consistent across languages by developing a cross-lingual depression detection framework based on the HuBERT model.

Previous research on SSL models demonstrated that speech representations from middle layers yield better results on emotion recognition across languages, especially when the model was trained on multilingual data [2]. These findings suggest that some depression-related acoustic patterns may also transcend language boundaries, enabling more generalizable models for early mental health screening [1].

The current study investigates which middle layers of HuBERT model generalize best across two different languages for the task of depression detection. HuBERT is trained for binary classification (depressed vs. non-depressed) on two datasets of depression speech: English (DAIC-WOZ) and Mandarin Chinese Multimodal Depression Corpus. HuBERT is fine-tuned using both datasets separately. Evaluation of the model performance is done on English, Mandarin, and mixed-language speech segments to investigate the efficiency of cross-lingual transfer. In particular, the work will assess whether a model trained on English can successfully detect depression in Mandarin speech and vice versa.

The findings of this study contributes to the growing body of research on speech modeling for automatic depression detection, suggesting that cross-lingual transfer may be an efficient strategy in cases of data scarcity and low-resource language context. Future work will further explore how different HuBERT layers encode depression-relevant representations across languages and examine the impact of label types—such as self-reported scales versus clinician-administered assessments—on model performance and cross-lingual generalization.

Contents

1 Introduction

Depression is a common psychiatric disorder, with symptoms including persistent sadness, anhedonia, cognitive impairment, and physical symptoms including sleep and appetite disturbances, according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Depression is a global challenge and significantly increases disability and suicide risk. Clinical assessment of depression typically relies on structured interviews and standardized assessment scales, such as the Patient Health Questionnaire-9 (PHQ-9), the Hamilton Depression Rating Scale (HDRS), and the Beck Depression Inventory (BDI). While these instruments have been extensively validated, they are labor-intensive, require highly trained professionals, and rely heavily on subjective self-reports, limiting their scalability [3].

Early detection of depression is critical to improving treatment outcomes and reducing the burden on healthcare systems. In recent years, speech-based depression detection has emerged as a promising non-invasive approach, as speech can naturally encode emotional and cognitive states. Studies have shown that depression affects prosodic and acoustic features[4].

Meanwhile, advances in self-supervised learning of speech representations (SSL) have revolutionized the way models learn from unlabeled data. SSL models like HuBERT [5] can learn high-level acoustic representations by predicting masked speech segments, thereby performing well in low-resource environments and downstream tasks such as emotion recognition and depression detection [2, 1]. Notably, the intermediate layers of HuBERT (e.g., layers 6-9) have been shown to encode transferable paralinguistic information across languages [2], which is particularly important for speech modeling related to depression.

By exploring the potential of using self-supervised speech models for language-independent depression detection, this study contributes to building more inclusive, accessible, and scalable mental health assessment tools for linguistically diverse populations.

1.1 Research Questions and Hypotheses

Based on the literature review, this study aims to explore whether HuBERT-based speech vectors can encode acoustic cues related to depression and make them generalizable across typologically different languages, specifically English (a non-tonal language) and Mandarin (a tonal language). The primary goal of this study is to evaluate whether cross-lingual and mixed-language training can improve model performance and representation robustness.

The main research question is: To what extent does mixed-language training enhance the generalizability of HuBERT-based speech embeddings for depression detection across English and Mandarin, as measured by cross-lingual classification performance (F1-score, accuracy, and ROC-AUC)?

This question is addressed through the following sub-questions:

Sub-question 1: How do HuBERT embeddings from different middle layers (Layers 6–9) affect cross-lingual depression detection performance?

Sub-question 2: To what extent can a model trained on one language (English or Mandarin) generalize to detecting depression in the other, as measured by F1-score, Accuracy, and ROC-AUC?

Sub-question 3: Does mixed-language training improve cross-lingual depression detection performance compared to monolingual training?

Grounded in findings from [2] and [1], the following hypotheses are proposed:

H1: Mixed-language training improves cross-lingual depression detection performance compared to monolingual models, particularly when using HuBERT middle-layer embeddings.

H2: HuBERT middle layers (specifically Layers 6 and 7) provide more transferable and effective representations for depression detection across languages than deeper layers (e.g., Layers 8 and 9).

H3: Models trained on mixed-language datasets achieve higher F1-score, accuracy, and ROC-AUC in zero-shot cross-lingual depression detection than models trained on a single language.

These hypotheses are tested through a controlled experimental design that evaluates HuBERT embeddings under different training conditions (monolingual vs. mixed-language) and at different levels. The next section details the methodology used to implement and evaluate this research.

2 Literature Review

2.1 Depression and Depression Speech

Depression affects not only thoughts and emotions. It also changes how people speak. Speech may carry useful signals for detecting depression. Some of the earliest findings are discussed by Cummins et al. [3]. They identified common prosodic features in depressed speech, such as reduced pitch range, slower rate, longer pauses, and monotone voice. These features may reflect emotional flatness or psychomotor slowing.

Other studies have confirmed these patterns but also pointed out important differences. It was found that in depression speech, vowel duration and pause patterns changed [6]. However, pitch did not always vary. This suggests that pitch may not work as a reliable marker across all populations. Kappen et al. [4] offered a broader review. They confirmed that prosodic cues are affected by depression. Yet, they also highlighted major differences across studies. These include inconsistent clinical labels, varied recording settings, and limited language diversity. Many studies are restricted to English-speaking participants.

High accuracy was achieved by Faurholt-Jepsen et al. [7] to identify depression from speech. They used pitch, speech rate, and loudness in the model. Data was however gathered from a lab-based German-speaking population. The results may not generalize to natural speech or to different languages.

These findings show that speech has informative cues about depression presence. But current evidence is not conclusive. The majority of work uses hand-engineered acoustic characteristics. They also work on small data sets and are not cross-linguistically validated. This study builds on those findings, following a different route. The study uses self-supervised learning (SSL) embeddings for learning depression-related acoustic cues.

2.2 Depression Speech with Self-Supervised Learning

The previous section reviewed how depression can alter acoustic speech features such as pitch variability, speech rate, and pause patterns. These patterns are subtle and variable across speakers and languages, making them difficult to model consistently. Early studies relied on hand-crafted acoustic features such as MFCCs, jitter, shimmer, and pitch-based measures [8, 6]. These were typically used with traditional classifiers like support vector machines or random forests. While interpretable, these systems required extensive preprocessing and tended to be brittle across languages, speakers, or recording conditions.

To address these challenges, deep learning approaches introduced data-driven methods for modeling speech signals. Recurrent neural networks, such as bidirectional LSTMs, were applied to predict depression severity from clinical interview speech [9]. These models captured temporal dynamics more effectively than static features. However, they also required large labeled datasets and remained sensitive to linguistic variation.

Self-supervised learning (SSL) has recently emerged as a powerful alternative. Instead of relying on annotated data or task-specific features, SSL models are trained to reconstruct or discriminate masked segments of raw speech. This process yields general-purpose representations that can be fine-tuned for a range of downstream tasks. Models such as wav2vec 2.0 and HuBERT exemplify this trend [10, 5]. In depression detection, SSL-based representations have outperformed traditional MFCCs, offering improved robustness and transferability [11, 12].

However, not all SSL models are equally well-suited for capturing paralinguistic information. For instance, wav2vec 2.0 is trained to predict fine-grained latent features optimized for phoneme-level recognition [13]. This design benefits automatic speech recognition (ASR) but may suppress prosodic and affective cues crucial for identifying depression. In contrast, alternative architectures focus on higher-level speech structure. These include models that use discrete unit prediction to encourage representation of rhythm, timing, and speech boundaries.

HuBERT exemplifies this shift. It learns to predict masked speech units derived from unsupervised clustering, which makes it more sensitive to suprasegmental features than phoneme-centric models [5]. Furthermore, its architecture supports layer-wise probing, allowing researchers to explore which layers encode prosodic information more effectively. Recent studies have compared HuBERT with other SSL models and found that intermediate layers often generalize better in affective tasks, especially in cross-lingual depression detection scenarios [1]. These characteristics make HuBERT a promising candidate for modeling the acoustic markers of depression in multilingual settings.

The following section builds on this foundation by examining how HuBERT's internal layers encode depression-relevant cues, and whether these representations can generalize across languages such as English and Mandarin.

2.2.1 HuBERT Layer-wise Representations and Cross-Lingual Depression Modeling

HuBERT is a self-supervised speech model that learns acoustic representations by predicting masked units from quantized speech segments [5]. Its architecture includes a convolutional feature encoder followed by a 12-layer Transformer network. Each layer in this structure processes information at different levels of abstraction, ranging from low-level spectral patterns to high-level linguistic representations. This makes HuBERT suitable not only for phoneme recognition, but also for modeling complex paralinguistic features such as prosody and vocal variability.

Recent studies suggest that different HuBERT layers capture different types of speech information, and that intermediate layers are particularly effective for paralinguistic tasks. Maji et al. [1] conducted a layer-wise analysis of HuBERT embeddings in a depression detection task involving Bengali and English speech. They found that embeddings from the middle layers (Layers 6–9) led to the best performance, both in monolingual and cross-lingual settings. This finding aligns with previous emotion recognition studies, where middle layers were also found to capture affective cues more robustly than shallow or deep layers.

These findings are further supported by Han et al. [2], who examined SSL representations in cross-lingual emotion recognition. They compared the performance of different SSL models—including

HuBERT and wav2vec2—on emotion classification across languages. Their results showed that HuBERT's intermediate layers achieved higher accuracy and better generalization than deeper or final layers, especially when tested on mismatched language pairs. This suggests that middle layers retain transferable prosodic and affective information that is less tied to language-specific phonetic patterns.

Although large-scale evaluations such as the SUPERB benchmark [14] do not provide layer-wise results, they consistently report strong performance for HuBERT on paralinguistic tasks like speaker identification and emotion recognition. Together with more targeted probing studies, this suggests that HuBERT's internal representations—especially at intermediate depths—are well-suited for modeling non-linguistic acoustic cues relevant to depression detection.

These observations provide the foundation for the present study. Since depression-related prosodic features must generalize across languages, it is important to identify which layers encode such features most effectively. By comparing HuBERT embeddings from different layers on English and Mandarin speech, this study aims to determine whether middle layers indeed offer the best trade-off between language-specific and transferable acoustic representations.

However, the effectiveness of HuBERT embeddings in cross-lingual depression detection depends not only on the model architecture, but also on the characteristics of the languages involved. The next section discusses linguistic and cultural factors that may influence how depression is manifested in speech across English and Mandarin.

2.3 Linguistic Considerations: Prosody, Tone, and Cultural Expression

Prosodic features differ significantly across languages, and these differences can influence how depression manifests in speech. English is a stress-based language, where intonation carries emotional nuance and pragmatic meaning. In contrast, Mandarin Chinese is a tonal language. It uses pitch contours lexically to distinguish word meanings. This phonological difference affects how pitch variation can be interpreted. In tonal languages like Mandarin, pitch is tightly constrained by tone identity, which limits its availability as an emotional or prosodic marker [15]. This has implications for depression-related features such as monotonicity or flattened pitch, which are often reported in English but may be masked by tone structure in Mandarin [16].

Beyond phonology, cultural norms also shape how individuals express or suppress emotional distress vocally. In some cultural contexts, flat or withdrawn speech may be expected or socially reinforced during emotional suffering. Kirmayer [17] emphasized that cultural display rules affect both the perception and production of vocal emotion. This means that vocal cues commonly associated with depression in one language or culture may not appear the same in another. For instance, emotional expressiveness is often less overt in East Asian speech cultures compared to Western ones, particularly in formal or clinical contexts [18].

These phonological and cultural differences also raise challenges for cross-lingual modeling. Although models like HuBERT aim to learn language-agnostic speech representations, the structure of the input language can influence how acoustic features are encoded. As such, a model trained on English speech may learn different depression-related cues than one trained on Mandarin. The interaction between linguistic form and acoustic signal makes it necessary to evaluate whether representations are truly transferable across typologically distinct languages.

Taken together, previous work shows that speech contains valuable cues for detecting depression, particularly in prosodic and timing-related features. However, these features are not always consistent across studies or languages. HuBERT offers a promising alternative, and evidence suggests that its intermediate layers may generalize better across tasks and languages. Yet, this remains underexplored in the context of depression, especially between English and Mandarin. This study addresses that gap by conducting a layer-wise analysis of HuBERT embeddings, investigating whether depression-relevant representations are transferable across languages and which layers offer the most reliable cross-lingual generalization.

3 Methodology

This chapter describes how the experiment was designed and carried out. It builds directly on the findings and motivations from the literature review. Previous studies suggest that depression affects prosodic speech patterns. They also show that self-supervised models like HuBERT can capture these patterns, especially in their middle layers. This chapter outlines the datasets, processing steps, model setup, and evaluation strategy used to test those ideas.

3.1 Dataset Description

3.1.1 Extended DAIC-WOZ Corpus (E-DAIC)

An extended version of Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) database that contains English semi-clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The dataset includes 219 interviews between interviewers and patients, and each includes Patient Health Questionnaire (PHQ-8) scores and PHQ binary.

3.1.2 Chinese Mandarin Depression Corpus (CMDC)

The Chinese Mandarin Depression Corpus (CMDC) is a publicly available Mandarin speech dataset designed to support automatic depression screening. The corpus was introduced by [19] to address the scarcity of Mandarin-language resources for speech-based mental health research.

CMDC consists of semi-structured clinical interviews conducted with both depressed patients and healthy controls. The interviews were designed to elicit spontaneous and emotionally relevant speech, covering a range of topics commonly used in psychological assessment. Each recording includes PHQ-9 scores and corresponding depression diagnoses.

The dataset comprises audio recordings organized at the question level (Q1–Q7), with each recording capturing a participant's response to a specific interview question. In total, CMDC contains data from 116 participants [19].

In this study, a balanced subset of 52 participants was selected for experimentation, consisting of: 26 healthy controls (HC01–HC26), and 26 depressed patients (MDD01–MDD26).

The use of CMDC enables the investigation of cross-lingual depression detection in a low-resource language context, complementing the English-based E-DAIC dataset.

3.2 Core Methods and Models

The core of this study's approach lies in leveraging self-supervised learning (SSL) for speech, specifically through HuBERT [5], to extract acoustic representations relevant to depression. Unlike traditional feature engineering, HuBERT enables data-driven learning of latent features directly from raw waveforms. Previous research has demonstrated that middle layers of HuBERT (particularly Layers

6–9) encode paralinguistic signals that correlate with emotion and affective states [2]. This study therefore focuses on these layers to examine their utility in detecting depression across languages.

For classification, a logistic regression classifier is used. The architecture is intentionally kept simple to ensure that observed performance differences are attributable to the input representations rather than model complexity. Binary depression labels are used to train the classifier. Comparisons are made across three training configurations: (1) monolingual-English, (2) monolingual-Mandarin, and (3) mixed-language (balanced English + Mandarin).

3.2.1 Data preparation

To ensure consistency and compatibility with the HuBERT architecture [5], all audio data underwent a unified preprocessing pipeline. This pipeline was implemented using Python scripts specifically designed for segment-level processing of each dataset, and consisted of the following steps:

First, all speech recordings were downsampled to 16 kHz, as required by the HuBERT feature extractor. Next, a segmentation strategy was applied to convert full-length audio files into shorter utterance-level segments. For the E-DAIC corpus, this involved two stages: (1) extracting interviewer–participant utterances from transcript-aligned audio using start–end timestamps, and (2) further dividing each utterance using a 3-second sliding window with 50% overlap. For the CMDC corpus, since each audio file corresponded to a single question response, the segmentation was applied directly on each Q-level recording using the same 3-second sliding window strategy with 50% overlap. No additional denoising or silence removal was applied.

Following segmentation, each resulting audio segment inherited the binary depression label of its parent unit: for E-DAIC, this was the label assigned to the original utterance; for CMDC, it was derived from the PHQ-9-based diagnostic label of the speaker.

In addition, all stereo recordings were converted to mono, both to reduce computational cost and to ensure consistent tensor shapes during HuBERT feature extraction. The preprocessed segments were then split into training (60%), validation (20%), and test (20%) subsets for each experimental condition.

Lastly, to ensure fair comparison between the English and Mandarin models and enable meaningful mixed-language training, data balancing was applied. While segmentation of E-DAIC produced approximately 31,296 segments, the CMDC set yielded 7,712. To match the scale of the Mandarin dataset, the E-DAIC training set was randomly downsampled to produce a balanced training set of 7,712 English segments, maintaining the same 60/20/20 distribution across splits. This ensured that model performance would not be skewed by differences in dataset size across experimental conditions.

Due to time constraints, this study adopts a clean, controlled experimental setup. Logistic regression was deliberately chosen over more complex models such as neural networks or support vector machines. This design isolates the representational differences across HuBERT layers and languages without introducing model biases or tuning artifacts. While advanced classifiers may improve raw performance, they risk confounding the analysis of how well HuBERT's internal representations encode depression-relevant cues.

3.2.2 HuBERT Feature Extraction

Following segmentation, feature extraction was performed using the HuBERT Base model pretrained on the LS-960 subset of LibriSpeech [5]. All audio segments were downsampled to 16 kHz, converted to mono, and normalized to ensure consistency with the model's training conditions. Hidden state representations were extracted from Layers 6 through 9, based on prior findings that these intermediate layers capture paralinguistic features transferable across languages [2]. For each segment, mean pooling was applied over the temporal dimension to produce a 768-dimensional embedding. Mean pooling was chosen because it simplifies temporal variation while preserving the average acoustic profile of each utterance. This helps retain prosodic cues such as pitch level and intensity range, which are relevant for detecting depression but do not require fine-grained temporal resolution. In addition, previous work [1] found that mean pooling outperformed max pooling in downstream depression classification tasks, particularly when using HuBERT embeddings. This reinforces its suitability for layer-wise benchmarking in this study.

3.2.3 Logistic Regression Classification Model

To evaluate whether these embeddings encode depression-relevant cues, logistic regression was selected as the baseline classifier. This model has been widely adopted in recent speech-based mental health research due to its simplicity and interpretability [1]. By using a linear classifier, the study ensures that observed differences in performance reflect the representational quality of the embeddings, rather than the capacity of a complex model.

Three experimental conditions were defined based on training language configuration: (1) monolingual-English, (2) monolingual-Mandarin, and (3) mixed-language (English + Mandarin).

3.3 Technical Framework

To ensure consistency and fairness across experiments, a standardized technical framework was designed to support utterance-level preprocessing, HuBERT-based feature extraction, and cross-lingual classification. All steps in the pipeline adhered to uniform segmentation criteria, speaker-level dataset splitting, and model evaluation procedures.

The E-DAIC dataset was first processed to extract participant utterances using metadata-aligned segmentation. Utterances were labeled with binary depression annotations and organized into a metadata table. To address class imbalance, a downsampling strategy was applied to the majority class, producing a balanced training set in accordance with best practices in speech-based depression detection [20].

Next, both E-DAIC and CMDC underwent fixed-length segmentation using a 3-second sliding window with 50% overlap. This windowing strategy ensured consistency across datasets and was chosen

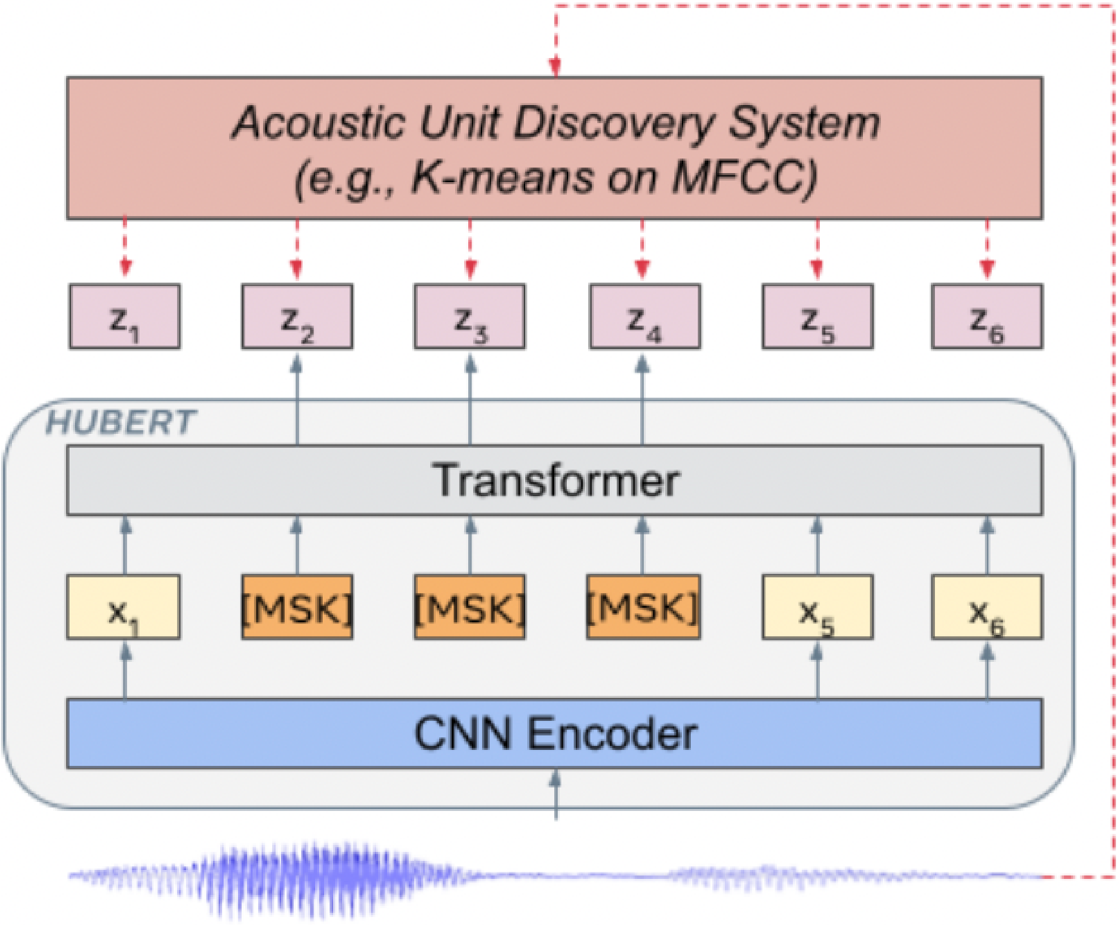


Figure 1: The overview of the HuBERT model architecture. Reprinted from [5].

to align with the HuBERT model's pretraining input format. The segmented data were then split into training, validation, and test sets in a 60/20/20 ratio at the speaker level, minimizing the risk of speaker overlap across subsets.

To construct a balanced mixed-language condition, equivalent quantities of English and Mandarin utterances were sampled. The mixed-language set was created by matching the number of Mandarin segments and drawing a random, speaker-balanced subset of English utterances. This procedure ensured that the MIX dataset included equal contributions from each language and could serve as a fair training condition for assessing Hypothesis **H1**.

Feature extraction was performed using the HuBERT Base model [5]. All segments were resampled to 16 kHz and normalized. Hidden states were extracted from Layers 6 through 9, and mean pooling was applied over time to generate fixed-length embeddings. These embeddings served as input to a logistic regression classifier.

For evaluation, models were trained on monolingual and mixed-language datasets and tested both

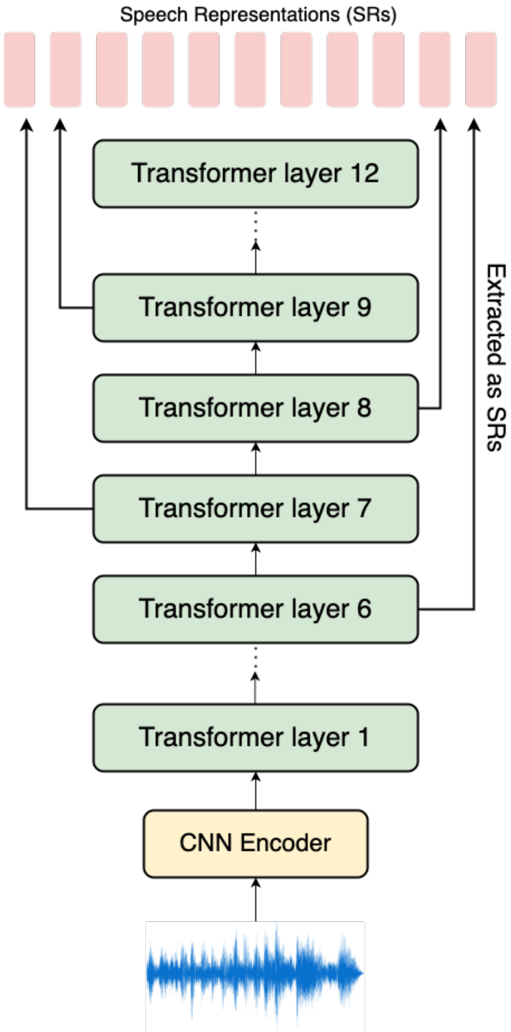


Figure 2: HuBERT architecture and layer-wise feature extraction. Speech representations (SRs) are extracted from layers 6–9 and used for downstream depression classification.

within-language and cross-lingually. Consistent with the literature, key metrics included F1 score, accuracy, and ROC-AUC [2, 1]. Models were trained with balanced class weighting to mitigate label skew, and all data preprocessing and model evaluation steps used fixed random seeds to ensure reproducibility.

3.4 Evaluation Methodology

The evaluation methodology was designed to assess how well HuBERT-based embeddings support cross-lingual generalization in depression detection, under different training conditions and layer configurations. Drawing on recent studies in speech emotion recognition and SSL-based clinical modeling [2, 1], the study adopts a strict zero-shot transfer framework, where classifiers trained on one language are directly tested on a different language without fine-tuning.

Four experimental conditions are defined:

EN → ZH: trained on English (E-DAIC), tested on Mandarin (CMDC)
ZH → EN: trained on Mandarin (CMDC), tested on English (E-DAIC)
MIX → ZH: trained on a balanced English-Mandarin dataset, tested on CMDC
MIX → EN: trained on the same balanced set, tested on E-DAIC

This structure allows examination of asymmetric transferability (e.g., EN→ZH vs. ZH→EN) and whether training on multilingual data improves generalization, particularly in relation to H1.

Each classifier is trained using logistic regression with balanced class weighting to account for any residual label imbalance. Evaluation is conducted independently for each of the four HuBERT layers (6–9), enabling layer-wise comparison of model performance. Metrics include F1 score (primary), accuracy, and ROC-AUC, which together capture class-sensitive performance, overall correctness, and decision separability.

This layer-wise analysis tests H2, by identifying which embedding layers contribute most to generalization. The results also allow exploration of whether multilingual exposure (H3) supports more robust depression detection across phonologically distinct languages.

3.5 Ethics and Research Integrity

This chapter discusses the ethical principles that guided the design and implementation of this study, and outlines the steps taken to ensure data privacy, scientific transparency, and research integrity.

3.5.1 Data Ethics and Privacy

All speech data used in this study were obtained from publicly available deidentified datasets: The Extended DAIC-WOZ Corpus (E-DAIC) [21] is released under controlled access and has been fully anonymized to remove personally identifying information. The Chinese Mandarin Depression Corpus (CMDC) [19] has also been similarly deidentified and is available for academic research purposes only.

No attempt was made to re-identify the participants, nor to infer personal characteristics beyond the scope of the provided labels (depressive states). Only speech data and the associated depression labels were processed, following the principles of data minimization and purpose limitation. No other metadata such as age, gender, or socio-demographics were used.

As this study involved secondary analysis of an existing dataset, no new data were collected from human participants and therefore no additional ethical approval was required.

3.5.2 FAIR Principles Implementation

The study was designed with reference to the FAIR principles, which advocate for making research data and methods Findable, Accessible, Interoperable, and Reusable:

Findable: All intermediate metadata tables (e.g., segment metadata, segment assignments) were carefully versioned and documented. **Accessible:** Code and processing scripts were structured to allow easy replication by other researchers with access to the same dataset. **Interoperable:** Standard data formats (e.g., .wav, .csv, .npy) were used throughout this study. **Reusable:** The entire processing and modeling pipeline was implemented using widely adopted open source libraries (Transformers, Torchaudio, scikit-learn), ensuring that these methods are widely applicable to similar speech research.

3.5.3 Open Science Practices

The full code base of this project, including all preprocessing, feature extraction, and classification scripts, is versioned using Git and will be publicly available through GitHub after the paper is submitted.

For further transparency, the exact versions of all software dependencies (Python 3.8, HuggingFace Transformers, scikit-learn, Torchaudio) are documented in this paper. This ensures that future researchers can repeat experiments with consistent software environments.

3.5.4 Bias and Fairness

Although the datasets used were balanced at the label level (depressed vs. non-depressed), potential biases still exist due to differences in participant demographics, recording conditions, and cultural-linguistic factors: The E-DAIC dataset was collected in the United States, while the CMDC dataset was collected in China. Differences in cultural representations of depression may affect linguistic features [1]. This study did not control for gender and age distribution because these metadata were not fully available in each dataset. No explicit debiasing techniques were used other than ensuring label balance and using `class_weight='balanced'` in classification.

3.5.5 Environmental Impact

Both training and inference were performed on Habrok GPU clusters, which use energy-efficient NVIDIA A100 GPUs. The computational impact of this research was intentionally kept modest: The model architecture (HuBERT Base) is a pretrained model; no large-scale pretraining was performed. Only a logistic regression classifier was trained, with minimal hyperparameter tuning. The processing scripts were designed for efficient batch processing and to minimize idle compute time.

In this regard, the carbon footprint of this research is much lower than typical deep learning pipelines involving full model fine-tuning or training large Transformer models from scratch.

3.5.6 Reproducibility and Replicability

This study aims to be fully reproducible: All scripts are version controlled. All random seeds are fixed (e.g., for data splitting, sampling, and classifier initialization). Preprocessing parameters (e.g., sliding window size, overlap, downsampling ratio) are fully documented. Feature extraction and evaluation outputs are systematically saved to ensure transparency.

In addition, the study supports replicability: any researcher with access to the E-DAIC and CMDC datasets can reproduce the results using the provided code base.

This study was designed and conducted with the principles of responsible data use, scientific transparency, and open sharing in mind. Although cross-lingual depression testing remains a challenging and sensitive area of research, this study contributes to the field in a manner that meets the highest standards of research ethics and integrity.

3.6 Demonstrator

This demonstrator accompanies the MSc thesis titled "Layer-wise Cross-Lingual Depression Detection from Speech: A HuBERT-Based Study on English and Mandarin." It provides a concise and reproducible implementation of the experimental pipeline used in the study. The goal is to show how self-supervised speech representations, specifically HuBERT embeddings, can be used to detect depression in a cross-lingual setting. By focusing on a layer-wise analysis across multiple training configurations, this demonstrator highlights how acoustic cues related to depression may transfer between distinct languages.

Features:
1. Cross-lingual depression detection using speech from two languages: English (CMDC) and Mandarin (E-DAIC)
2. Layer-wise analysis of HuBERT embeddings, focusing on Layers 6–9 to evaluate their encoding of depression-relevant acoustic cues
3. Used a simple logistic regression classifier to isolate feature performance
4. Balanced multilingual training through segment-level data augmentation and downsampling
5. Evaluation across three training settings: English-only, Mandarin-only, and mixed-language, enabling insight into generalization and transferability
6. Results are presented using macro F1, accuracy, and ROC-AUC, with clear visualization of layer-wise trends

Technical Implementation:
1. Audio preprocessing: utterance-level segmentation using a 3-second sliding window with 50% overlap
2. HuBERT feature extraction: based on the HuBERT Base model, using layers 6–9
3. Mean pooling applied to obtain fixed-size utterance embeddings (768-dimensional)
4. Logistic regression classifier implemented with scikit-learn, using stratified train/validation/test splits

5. Reproducibility ensured through fixed seeds, public code, and full documentation

GitHub repository:

`https://github.com/querodormir/HuBERT_Depression_Detection`

4 Experimental Setup

The experiments in this study aim to systematically and controlled evaluate the cross-lingual generalization ability of HuBERT-based speech vectors for depression detection. The core goal is to explore the generalization ability of models trained in one language to another language, and whether mixed-language training can improve generalization performance.

To this end, the experimental setup includes carefully controlled data preparation, data segmentation, and data balancing procedures (see Chapter 3), as well as consistent feature extraction and classification processes across all experimental conditions.

The following sections detail the data preparation strategy, data segmentation procedures, construction of development and test subsets, and the design of two core experiments that aim to address the research questions and hypotheses of this study.

4.1 Experiment 1: Monolingual-to-Cross-Lingual Transfer

Two cross-lingual experiments were designed in this study. The goal of the first experiment was to evaluate to what extent a classifier trained only on monolingual data (English or Mandarin) can generalize to depression detection in an unknown target language. This setting is consistent with previous studies on cross-lingual emotion recognition and depression modeling [2, 1].

Two independent classifiers were trained: one using only E-DAIC English features (from layers 6, 7, 8, or 9) and the other using only CMDC Mandarin features. Both classifiers were trained using logistic regression on their respective training sets, validated on a validation set of the matching language, and then tested on a held-out test set of the other language without any form of adjustment or fine-tuning.

This design simulates a real zero-resource environment, where annotated speech in the target language may not be available and the model must rely entirely on knowledge transferred from the relevant source language. This study also analyzed the performance differences of different layers to evaluate which HuBERT representations can better support transferable depression-related information.

4.2 Experiment 2: Mixed-Language Training for Cross-Lingual Transfer

In the second experiment, the focus shifted to testing whether mixed-language training, combining English and Mandarin segments, could improve cross-lingual generalization. To construct the training data, a balanced mixture of E-DAIC and CMDC segments was sampled (50% per language), totaling 7,710 segments, to match the number of available Mandarin utterances after segmentation and downsampling.

The mixed dataset was split into train, validation, and test sets using a fixed 60/20/20 ratio by speaker.

A logistic regression model was trained on the mixed train set, validated on the mixed dev set, and then tested independently on the monolingual test sets of both English and Mandarin. This allowed for direct comparison between single-source and mixed-source training conditions.

This experiment was designed to evaluate whether language diversity in training leads to more robust depression representations in self-supervised embeddings. It also enables comparison with prior multilingual SSL findings suggesting improved generalization from mixed-language data [2].

4.3 Hyperparameter Settings

To ensure comparability across conditions, all experiments were conducted using identical model and evaluation settings, consistent with prior literature [20, 1]. The following configuration was adopted for all logistic regression experiments:

- **Classifier:** LogisticRegression from scikit-learn
- **Class balance:** `class_weight='balanced'`
- **Maximum iterations:** 1000
- **Random seed:** `random_state=42`
- **Feature scaling:** StandardScaler (fit on train set, applied to all splits)
- **Evaluation metrics:** F1 score, accuracy, and ROC-AUC
- **Test protocol:** test set strictly held out during training and validation
- **Layer-wise configuration:** experiments repeated for each of HuBERT Layers 6–9

This configuration ensures reproducibility across experiments and eliminates variance introduced by hyperparameter tuning. Using a lightweight linear classifier (logistic regression) allows for clearer attribution of performance differences to representational factors in the embeddings rather than model architecture or capacity.

Together, these experiments form a comprehensive framework for evaluating the cross-lingual utility of self-supervised speech representations for depression detection. Experiment 1 provides a baseline assessment of zero-shot generalization capabilities for monolingual training. Experiment 2 explores whether exposure to multiple languages during training leads to more transferable representations.

By coordinating data processing, segmentation, feature extraction, and evaluation across all conditions, this setup ensures that observed performance differences reflect differences in training data and choice of embedding layers.

5 Results

This section presents the performance results of the depression detection models across different training conditions and HuBERT layers. The focus is on evaluating how well HuBERT embeddings, extracted from layers 6 to 9, support both within-language classification and cross-lingual transfer.

Section 5.1 reports the results of models trained and tested on the same language, including English (E-DAIC), Mandarin (CMDC), and the combined MIX condition. Section 5.2 presents the results of zero-shot and mixed-language transfer experiments, in which the training and test languages differ. Key performance metrics include F1 score, accuracy, and ROC-AUC.

5.1 Within-Language Performance (Monolingual Training and Testing)

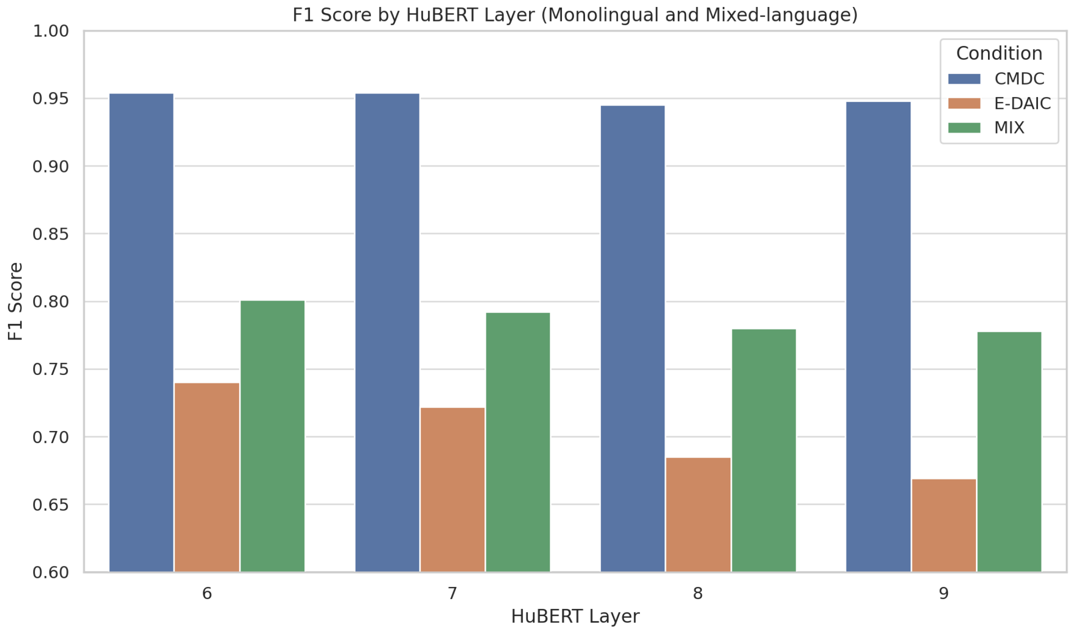


Figure 3: F1 scores across HuBERT layers for monolingual and mixed-language models.

This section reports the results of models trained and tested on the same language. The training conditions include CMDC (Mandarin), E-DAIC (English), and MIX. The models were evaluated across HuBERT layers 6 to 9. Figures 3 to 5 show the F1 scores, accuracy, and ROC-AUC for each setup.

CMDC models achieved the highest scores across all layers. The F1 score remained close to 0.95. Accuracy exceeded 95%, and ROC-AUC was consistently above 0.99. These results indicate strong within-language performance on Mandarin speech. The differences between layers were small.

In contrast, E-DAIC models had lower performance. The F1 score was 0.74 at Layer 6 and dropped to 0.67 at Layer 9. A similar trend appeared in the ROC-AUC scores. This decline suggests that deeper layers may carry less useful information for depression detection in English.

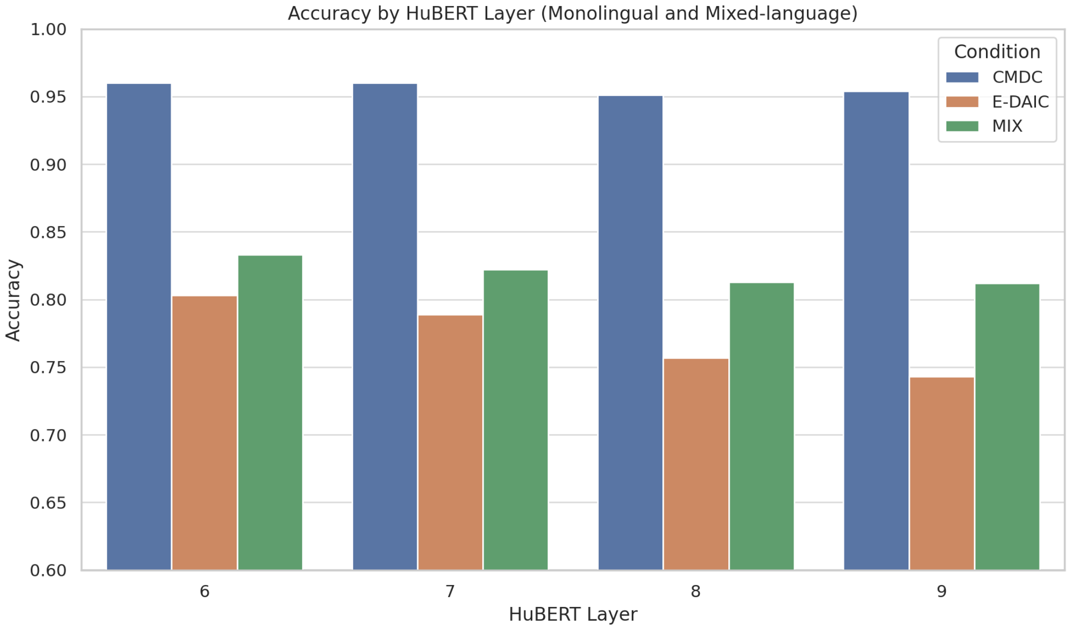


Figure 4: Accuracy by HuBERT layer for CMDC, E-DAIC, and MIX training conditions.

The MIX condition performed between CMDC and E-DAIC. The F1 score reached 0.80 at Layer 6 and stayed relatively stable across layers. Accuracy ranged from 81% to 83%. These results suggest that mixed-language training helps the model generalize better across different speech types, even without being language-specific.

Layer 6 and Layer 7 generally performed better under all training conditions. However, no significance tests were applied. The layer-wise trends are descriptive only and should not be interpreted as statistically confirmed.

5.2 Cross-Lingual Performance

This section presents results from zero-shot cross-lingual and mixed-language transfer experiments. Each model was trained on one language or a mixed-language dataset, and then tested on a different target language without additional fine-tuning. Performance was evaluated on English and Mandarin test sets across HuBERT layers 6 to 9. Figures 6 and 7 show the F1 scores and ROC-AUC values for all transfer conditions.

The **EN→ZH** condition produced weak results. The F1 scores were below 0.32 across all layers, with the highest value at Layer 8 (0.31). ROC-AUC peaked at 0.63. These results suggest that the English-trained model captured some depression-relevant patterns but failed to generalize well to Mandarin speech.

The **ZH→EN** condition performed even worse. F1 scores remained under 0.45, with the highest score at Layer 9. ROC-AUC was below 0.50 for all layers, indicating performance close to random chance. This asymmetry suggests that the Mandarin-trained model struggled to detect depression in

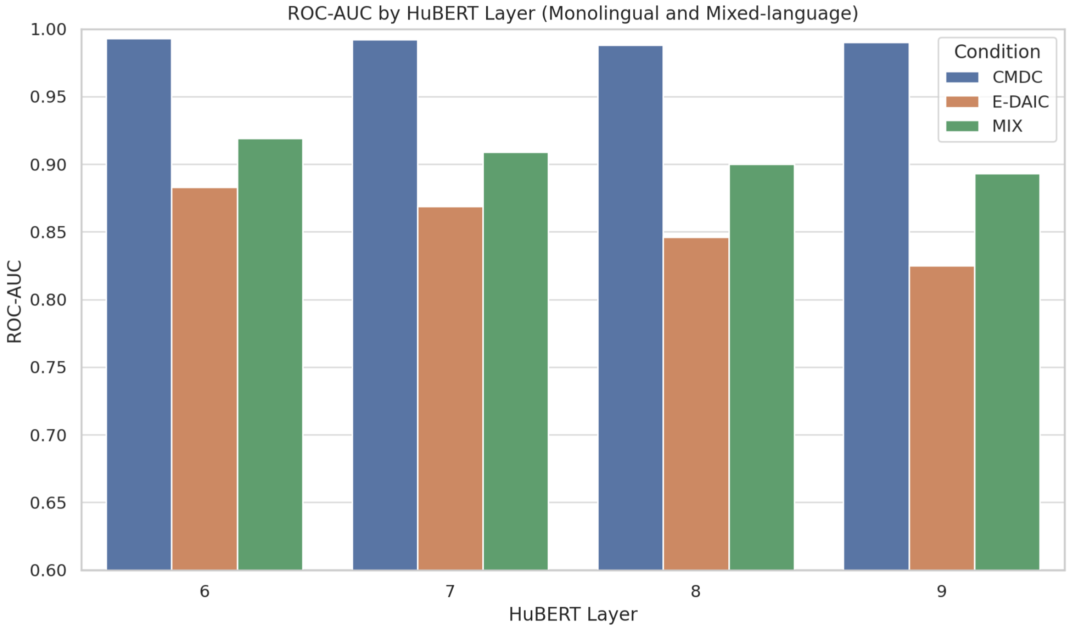


Figure 5: ROC-AUC across HuBERT layers under different within-language training setups.

English. However, this difference may also reflect dataset differences in recording quality, speech style, or diagnostic labels.

The **MIX→EN** condition showed stronger generalization. The best F1 score reached 0.65 at Layer 7. Accuracy and ROC-AUC also improved compared to ZH→EN. Similarly, the **MIX→ZH** condition achieved the best overall results. F1 scores were close to 0.92 at Layer 6 and 7, and ROC-AUC values exceeded 0.98.

Overall, the results suggest distinct performance trends between training conditions, target languages, and HuBERT layers. Within-language models performed best when trained on Mandarin. Cross-lingual transfer from mixed-language training yielded stronger generalization than training on a single language.

These observed patterns form the foundation for the next section. Section 6 will interpret these findings in the context of the study's hypotheses and previous research, and will also discuss potential explanations, limitations, and broader implications.

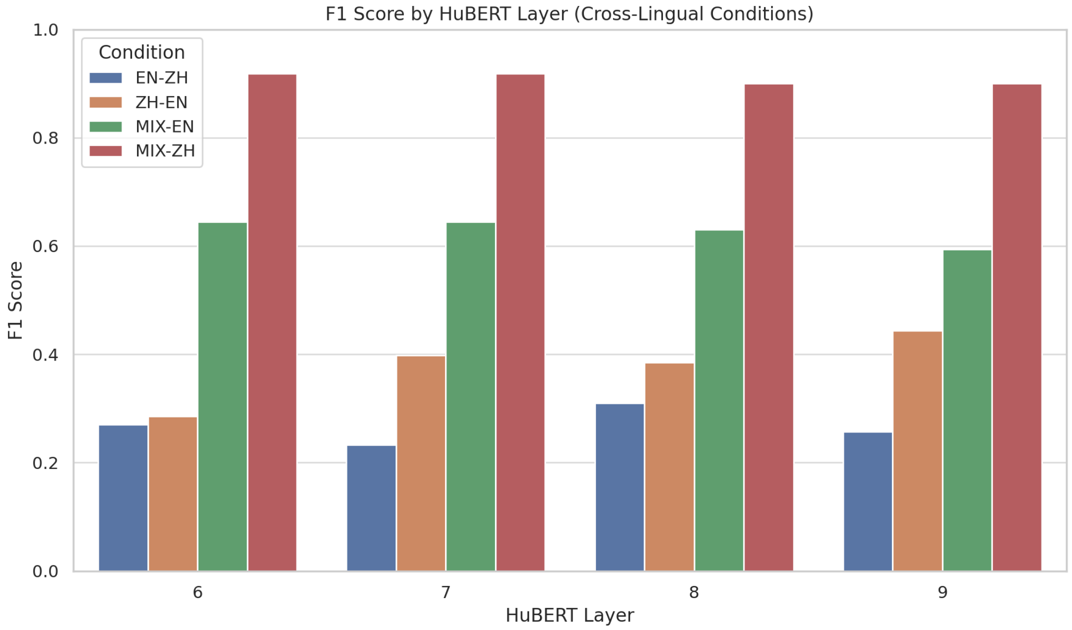


Figure 6: F1 scores across HuBERT layers for all cross-lingual transfer conditions.

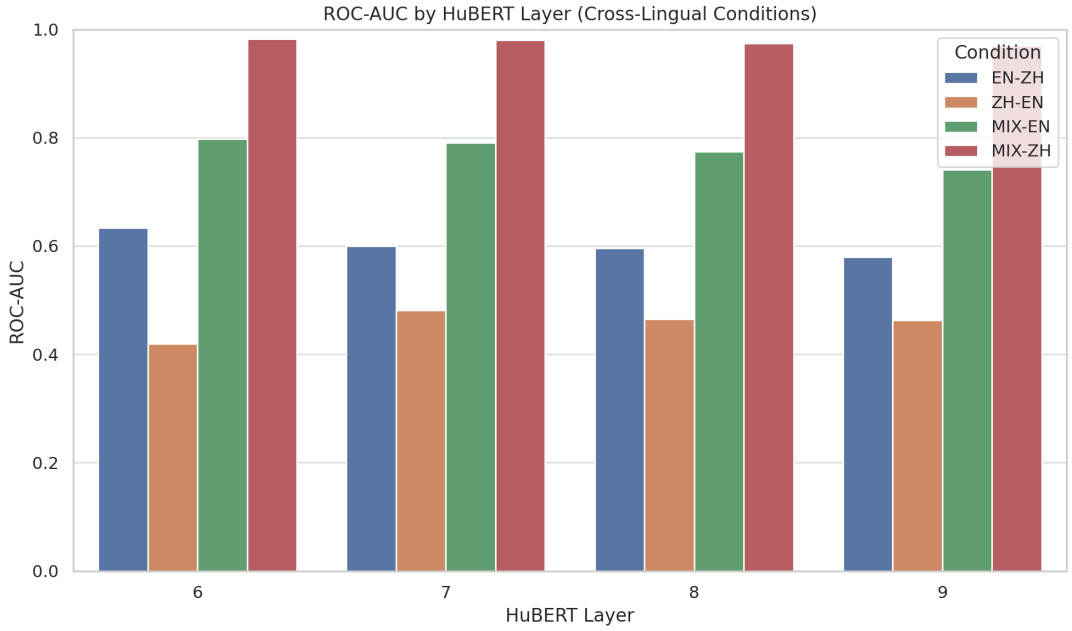


Figure 7: ROC-AUC across HuBERT layers for zero-shot and mixed-language transfer settings.

6 Discussion

This section discusses the experimental findings in relation to the research questions and hypotheses presented in Introduction. It aims to interpret the results by identifying patterns, connecting them to previous research, and explaining potential causes behind performance differences. The analysis also considers the limitations of the current study and reflects on the implications for future research in cross-lingual depression detection using self-supervised models.

6.1 Summary of Results

The results showed that Layer 6 and Layer 7 consistently yielded higher performance compared to deeper layers across all conditions. In within-language settings, CMDC-trained models performed best, followed by MIX, then E-DAIC. In cross-lingual tests, models trained on MIX data achieved better generalization than monolingual models. Notably, MIX→ZH performance nearly matched the CMDC baseline. In contrast, ZH→EN transfer was the weakest, with performance close to random guessing.

The following parts interpret these trends, examine their possible explanations, and evaluate how well they support the study's hypotheses.

6.2 Layer-wise Embedding Transferability

This subsection addresses sub-question 1, which examines how HuBERT embeddings from different middle layers (Layers 6 to 9) affect cross-lingual depression detection performance. Across all experimental conditions, Layers 6 and 7 consistently produced better results than deeper layers. This trend was observed in within-language models, mixed-language models, and zero-shot transfer settings.

The findings align with previous work suggesting that HuBERT's middle layers encode more transferable acoustic information [1]. Previous results showed that intermediate layers retain prosodic and affective cues, while deeper layers capture more language-specific or lexical features. Han et al. [2] had similar observations in emotion recognition. In their study, HuBERT's middle layers generalized better than final layers when tested across mismatched languages.

The current results support these interpretations. Models using embeddings from Layer 6 or 7 achieved the highest F1 scores and ROC-AUC values in nearly all conditions. In contrast, Layer 9 often showed degraded performance, particularly in cross-lingual settings. This suggests that deeper layers may contain more language-specific structure, which reduces transferability across languages.

These observations confirm H2. Intermediate layers offer a better trade-off between capturing depression-relevant prosodic patterns and maintaining generalization across tasks and languages.

6.3 Cross-Lingual Transferability of Monolingual Models

This subsection discusses sub-question 2, which asks whether models trained on a single language can generalize to a different language. In the current study, two zero-shot cross-lingual setups were tested: EN→ZH and ZH→EN. The results showed that both directions yielded poor performance. The EN→ZH model achieved an F1 score below 0.32 across all layers. The ZH→EN model performed even worse, with ROC-AUC values below 0.50 and F1 scores approaching random.

These findings suggest that monolingual models do not generalize well to a different target language. However, this limitation does not necessarily reflect inherent linguistic barriers. Instead, dataset-level differences likely played a larger role. For example, E-DAIC consists of spontaneous, open-domain speech with disfluencies, hesitations, and a wide range of speaking styles. In contrast, CMDC is a clean, structured dataset with controlled microphone settings and consistent speaker prompts. Models trained on CMDC may have overfit to these regularities and failed to adapt to the variability of English clinical speech.

Beyond dataset design, phonological and cultural factors may also contribute to transfer asymmetry. Mandarin is a tonal language, where pitch is tightly constrained by tone identity. In such cases, prosodic cues like monotonicity or pitch flattening—which are often used to signal depression in English—may be less reliable or even masked [15]. Cultural norms around emotional expression may further suppress vocal cues in Mandarin, particularly in formal or clinical contexts [17]. As a result, models trained on Mandarin data may encode a different set of acoustic depression markers than those trained on English, reducing cross-lingual compatibility.

Previous studies have also reported similar asymmetries in cross-lingual emotion recognition. Models trained on Mandarin generalized poorly to English, particularly when tested on highly expressive or prosodically diverse data [2]. This suggests that even if the model learns depression-relevant cues in one language, these cues may not transfer if their acoustic realizations differ too much between source and target datasets.

Overall, these results partially support H3. Monolingual models struggled to generalize across languages. The poor transfer performance highlights the need to account for dataset difference and acoustic mismatch in future cross-lingual depression detection systems.

6.4 Mixed-Language Training Improves Generalization

This part addresses sub-question 3, which asks whether mixed-language training improves cross-lingual depression detection performance. The results showed that models trained on the MIX dataset outperformed their monolingual counterparts in most cross-lingual settings. In particular, the MIX→ZH model achieved F1 scores above 0.91, nearly matching the CMDC monolingual baseline. MIX→EN also performed better than both EN→ZH and ZH→EN, with the highest F1 score reaching 0.65.

These results support H1 and partially support H3. Mixed-language training appeared to enhance the model's ability to generalize to unseen test languages. This improvement may be attributed to increased acoustic diversity in the training data. Exposure to multiple phonetic and prosodic systems may help the model learn more robust and language-agnostic depression indicators. This aligns with findings from Han et al. [2], who observed that training on multilingual data improved emotion recognition across languages.

The MIX training setup may have helped the model adapt to varied speech conditions. By learning from both Mandarin and English data, the model became less sensitive to the structural patterns of a single language. This could explain why MIX→ZH performance was nearly as high as the CMDC monolingual baseline, even though the model was not trained exclusively on Mandarin.

However, these findings remain descriptive. No significance testing was applied, and the results may be influenced by dataset difference. Nonetheless, the performance gap between MIX and monolingual transfer models is consistent across layers and metrics.

Overall, the results highlight the potential of mixed-language training to support cross-lingual depression detection without requiring language-specific fine-tuning.

6.5 Model Limitations

Beyond accuracy and F1 score, it is important to consider how and why the model may fail. This section discusses several key limitations observed in the current experiments. These include error tendencies, dataset constraints, and potential risks related to generalization and fairness. An inspection of confusion matrices showed that most models produced more false negatives than false positives. This means that depressed individuals were more likely to be misclassified as non-depressed. Such errors are especially concerning in clinical applications, where the cost of missing a diagnosis is higher than that of a false alarm.

Several factors may have contributed to this bias. First, training data may not fully reflect the variability of depressive speech. For example, the E-DAIC dataset[21] contains speech from a narrow demographic group, and CMDC[19] includes scripted responses with limited emotional range. These limitations reduce the model's exposure to diverse depression markers by gender, age, dialect, and speaking style.

Second, the model may rely too heavily on prosodic patterns that do not generalize across speakers or cultures. Depression manifests differently in speech across individuals and languages. Without explicit modeling of speaker or cultural context, the model may overfit to specific traits such as speaking rate, pitch contour, or intensity, rather than to depression itself.

These limitations highlight the need for fairness-aware evaluation in future work. While overall metrics may appear strong, performance should be assessed across different subgroups. Future studies could explore model calibration, subgroup performance breakdowns, or post-hoc error analysis to uncover hidden biases. These steps are essential to ensure that speech-based depression detection systems are safe, inclusive, and clinically reliable.

6.6 Summary

Section 6 discussed the experimental findings in relation to the research questions and hypotheses. It examined how model performance varied across HuBERT layers, training conditions, and test languages. Several consistent patterns emerged.

First, middle-layer embeddings from HuBERT (Layers 6 and 7) consistently outperformed deeper layers. This supports H2 and answers sub-question 1. These layers appear to retain prosodic and depression-relevant cues while avoiding overfitting to language-specific features.

Second, monolingual models failed to generalize across languages. Both EN→ZH and ZH→EN models yielded low F1 scores and poor AUC. These results partially support H3 and respond to sub-question 2. The asymmetry between directions may be explained by differences in dataset complexity and acoustic variability, rather than by linguistic distance.

Third, mixed-language training improved cross-lingual generalization. The MIX→ZH model achieved performance comparable to the monolingual Mandarin baseline. MIX→EN also outperformed other transfer models. These findings support H1 and further support H3. Sub-question 3 is thus answered affirmatively: training on diverse input conditions can enhance zero-shot detection in new languages.

The results presented here are descriptive. No statistical tests were applied, and comparisons across layers or models should be interpreted with caution. Nonetheless, the consistency of the observed trends across metrics and tasks suggests that mixed-language training and middle-layer embeddings are promising directions for future research.

7 Conclusions

This study explored whether HuBERT-based speech embeddings can be used for detecting depression across languages, with a particular focus on Mandarin and English. The study examined how different embedding layers and training strategies affect classification performance, using models trained on monolingual and mixed-language data. The experiments were designed to answer three research questions concerning embedding depth, zero-shot generalization, and the effectiveness of multilingual training.

Building on a comparative evaluation of multiple training conditions and HuBERT layers, the results provide new insights into how self-supervised models encode depression-related acoustic patterns. While the models showed strong within-language performance, especially on Mandarin data, cross-lingual generalization remained challenging. However, mixed-language training offered a promising strategy for bridging this gap.

7.1 Limitations

Limitations concern model design, data availability, and generalizability.

First, the study used a logistic regression classifier with fixed HuBERT embeddings. While this approach allows for interpretable comparisons across layers and conditions, it may not capture complex non-linear patterns in the data. More expressive classifiers, such as neural networks or ensemble models, could potentially improve detection accuracy.

Second, no statistical significance tests were applied. All performance comparisons were based on descriptive metrics from a single train-validation-test split. As a result, the reported trends should be interpreted cautiously and may not generalize to unseen samples or alternative data partitions.

Third, the datasets used in this study differ in size, recording conditions, and emotional style. CMDC consists of scripted, clean speech, while E-DAIC contains spontaneous clinical interviews with background noise and variable speech rates. These factors may confound cross-lingual comparisons. Moreover, labels were treated as binary (depressed vs. non-depressed), without considering the full spectrum of depression severity.

Finally, this study did not examine subgroup effects or fairness-related issues. The models were not evaluated across gender, age, or language proficiency subgroups. As depression is known to manifest differently across demographic and cultural contexts, future work should assess whether detection accuracy varies across speaker groups.

These limitations constrain the generalizability and robustness of the current findings. They also motivate several future research directions, outlined in the next subsection.

7.2 Future Work

Based on the limitations discussed above, several directions for future research are suggested.

First, future work can explore more advanced classification models. Neural networks, transformer-based classifiers, or ensemble methods could better capture non-linear and speaker-specific patterns in HuBERT embeddings. Comparing these models against logistic regression would provide insight into interpretability and performance.

Second, include statistical significance testing and cross-validation. Repeating experiments across multiple data splits would improve the robustness of the findings. This would also allow for hypothesis-driven comparisons between layers, training conditions, and languages.

Third, future studies can address dataset imbalance and complexity. Using more diverse and matched datasets would reduce confounding effects. In particular, training and testing on spontaneous speech in both languages could better reflect real-world conditions. Incorporating multi-level labels of depression severity, instead of binary classes, may also improve clinical relevance.

Another direction for future work involves error analysis of model predictions. This could include manual inspection of utterances that were consistently misclassified across multiple HuBERT layers or training conditions. Identifying linguistic or acoustic properties that lead to false positives or false negatives may provide insight into model limitations and dataset biases.

Finally, future research should examine fairness across speaker subgroups. Investigating model performance across gender, age, or cultural background would reveal potential biases. Including metadata and conducting subgroup analysis would be essential to build equitable and trustworthy systems.

These future directions aim to enhance both the methodological rigor and practical utility of cross-lingual depression detection.

7.3 Implications and Ethical Considerations

The findings of this study have broader implications for both speech-based mental health assessment and multilingual machine learning. First, the results suggest that mixed-language training can improve model robustness and generalizability. This is particularly relevant for applications in various languages, where building language-specific depression detectors may not be feasible. Using multilingual models could help extend access to mental health technologies across linguistic boundaries.

Second, the consistent performance of HuBERT's middle layers highlights the value of pre-trained self-supervised models in cross-lingual affective computing. Rather than designing new models for each language, researchers and developers may leverage language-agnostic representations to build scalable, cross-lingual systems.

However, these benefits also raise ethical concerns. Models that fail to generalize across speaker groups or cultural contexts may lead to unequal performance. For instance, speakers from cultures that suppress vocal expressivity may be more likely to be misclassified. If such systems are used in clinical or screening settings, these errors could reinforce existing health disparities.

In summary, this study offers promising directions for cross-lingual mental health modeling. At the same time, it emphasizes the need for fair and context-sensitive approaches before real-world deployment can be considered.

This study contributes to the growing body of research on speech-based mental health detection by exploring the cross-lingual behavior of HuBERT embeddings in depression classification. By conducting a layer-wise analysis across English and Mandarin, the study highlights the representational value of HuBERT's middle layers and underscores the challenges of transferring prosodic features across distinct languages. While limited in scope and statistical power, the findings helps for future multilingual affective speech research and point toward the importance of culturally sensitive modeling in mental health technology.

Bibliography

[1] B. Maji, R. Guha, A. Routray, S. Nasreen, and D. Majumdar, "Investigation of Layer-Wise Speech Representations in Self-Supervised Learning Models: A Cross-Lingual Study in Detecting Depression," in *Interspeech 2024*, 2024, pp. 3020–3024.

[2] Z. Han, T. Geng, H. Feng, J. Yuan, K. Richmond, and Y. Li, "Cross-lingual Speech Emotion Recognition: Humans vs. Self-Supervised Models," *arXiv preprint arXiv:2409.16920*, 2024.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[4] M. Kappen, M.-A. Vanderhasselt, and G. M. Slavich, "Speech as a promising biosignal in precision psychiatry," *Neuroscience & Biobehavioral Reviews*, vol. 148, p. 105121, 2023.

[5] 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," *arXiv preprint arXiv:2106.07447*, 2021.

[6] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," *PLOS ONE*, vol. 16, no. 4, p. e0248842, 2021.

[7] F. Menne, F. Dörr, J. Schräder, J. Tröger, U. Habel, A. König, and L. Wagels, "The voice of depression: Speech features as biomarkers for major depressive disorder," *BMC Psychiatry*, vol. 24, no. 1, p. 794, 2024.

[8] L.-C. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Automated assessment of depression from speech: A systematic review," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 251–270, 2022. [Online]. Available: https://doi.org/10.1109/TAFFC.2020.2973677

[9] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Interspeech 2018*, 2018, pp. 1716–1720. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2018/abstracts/0243.html

[10] 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. [Online]. Available: https://arxiv.org/abs/2006.11477

[11] P. Zhang, M. Wu, H. Dinkel, and K. Yu, "DEPA: Self-supervised audio embedding for depression detection," in *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, 2021, pp. 135–143.

[12] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," arXiv preprint arXiv:2104.03502, 2021. [Online]. Available: https://arxiv.org/abs/2104.03502

[13] Y. Li, Y. Mohamied, P. Bell, and C. Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," arXiv preprint arXiv:2210.02595, 2022. [Online]. Available: https://arxiv.org/abs/2210.02595

[14] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Qian, A. S. Subramanian, W.-C. Tseng, D.-R. Liu, Z. Huang, S. Dong, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-H. Weng, H. yi Lee, and J. Glass, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[15] Y. Wang and C.-C. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in mandarin chinese?" *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015. [Online]. Available: https://www.researchgate.net/publication/271536611_Does_restriction_of_pitch_variation_affect_the_perception_of_vocal_emotions_in_Mandarin_Chinese

[16] G. Peng, M. K. Chan, F.-M. Tsao, T.-R. Huang, and O. J. Tzeng, "Production and perception of tone and intonation in mandarin," *Journal of Neurolinguistics*, vol. 18, no. 6, pp. 437–456, 2005. [Online]. Available: https://doi.org/10.1016/j.jneuroling.2004.12.001

[17] L. J. Kirmayer, "Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment," *Journal of Clinical Psychiatry*, vol. 62, no. 13, pp. 22–28, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11434415/

[18] M. Starkey, "Cultural aspects of depressive experience and disorders," Grand Valley State University Undergraduate Research Journal, 2021. [Online]. Available: https://scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1081&context=orpc

[19] B. Zou, J. Han, Y. Wang, R. Liu, S. Zhao, L. Feng, X. Lyu, and H. Ma, "Semi-Structural Interview-Based Chinese Multimodal Depression Corpus Towards Automatic Preliminary Screening of Depressive Disorders," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2823–2838, 2023.

[20] Q. Zhao, H.-Z. Fan, Y.-L. Li, L. Liu, Y.-X. Wu, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, and S.-P. Tan, "Vocal Acoustic Features as Potential Biomarkers for Identifying/Diagnosing Depression: A Cross-Sectional Study," *Frontiers in Psychiatry*, vol. 13, p. 815678, 2022.

[21] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," n.d.

Appendices

A Declaration of AI Use

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used ChatGPT for the following purposes: sentence restructuring in sections, creating initial code documentation templates, summarizing background literature for preliminary review. All content was subsequently reviewed, verified, and substantially modified by me.

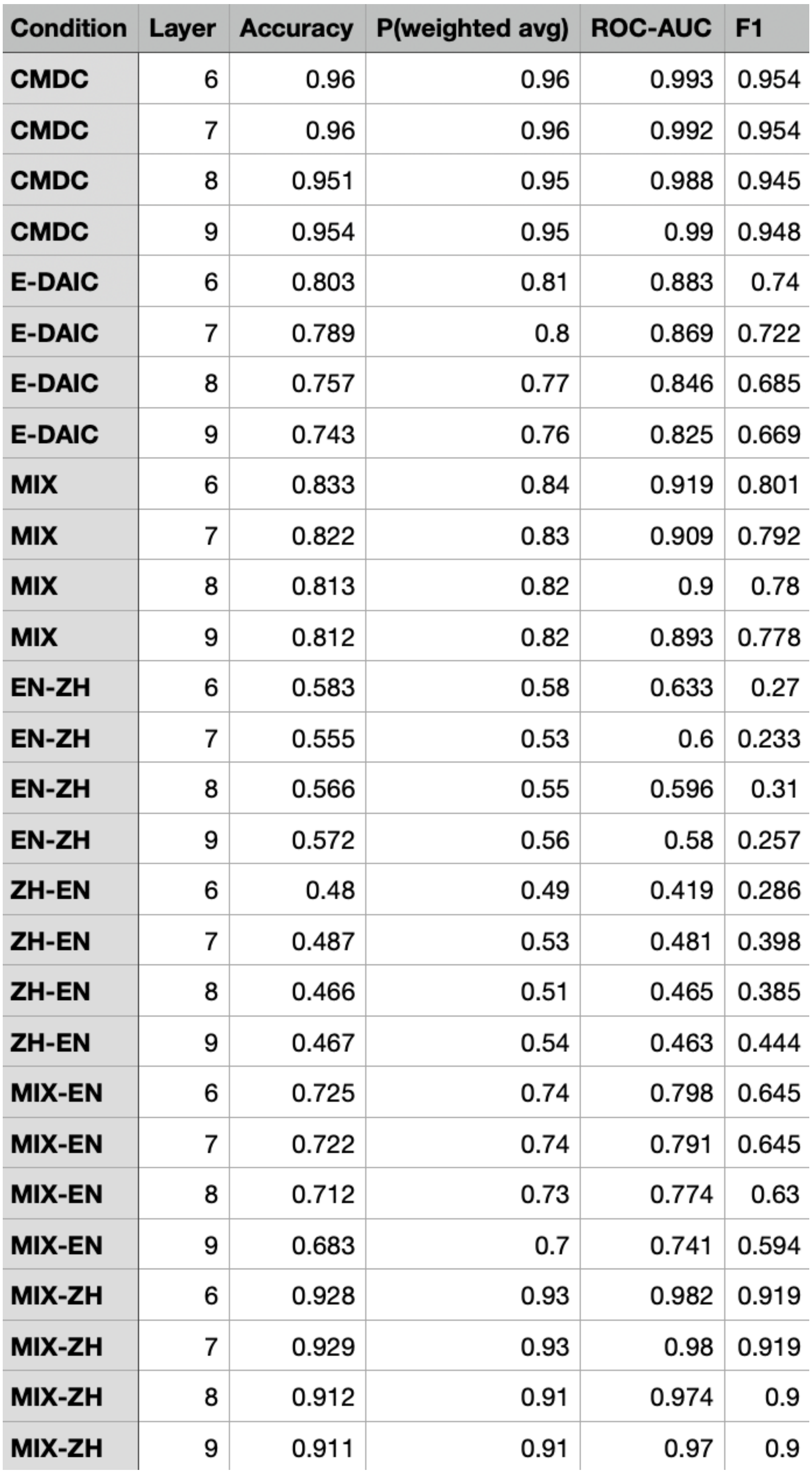
Hang Chen/2025-07-10

B Code Availability

The complete codebase used for all experiments in this study is available at the following GitHub repository:

`https://github.com/querodormir/HuBERT_Depression_Detection`

C Full Model Outputs



Condition	Layer	Accuracy	P(weighted avg)	ROC-AUC	F1
CMDC	6	0.96	0.96	0.993	0.954
CMDC	7	0.96	0.96	0.992	0.954
CMDC	8	0.951	0.95	0.988	0.945
CMDC	9	0.954	0.95	0.99	0.948
E-DAIC	6	0.803	0.81	0.883	0.74
E-DAIC	7	0.789	0.8	0.869	0.722
E-DAIC	8	0.757	0.77	0.846	0.685
E-DAIC	9	0.743	0.76	0.825	0.669
MIX	6	0.833	0.84	0.919	0.801
MIX	7	0.822	0.83	0.909	0.792
MIX	8	0.813	0.82	0.9	0.78
MIX	9	0.812	0.82	0.893	0.778
EN-ZH	6	0.583	0.58	0.633	0.27
EN-ZH	7	0.555	0.53	0.6	0.233
EN-ZH	8	0.566	0.55	0.596	0.31
EN-ZH	9	0.572	0.56	0.58	0.257
ZH-EN	6	0.48	0.49	0.419	0.286
ZH-EN	7	0.487	0.53	0.481	0.398
ZH-EN	8	0.466	0.51	0.465	0.385
ZH-EN	9	0.467	0.54	0.463	0.444
MIX-EN	6	0.725	0.74	0.798	0.645
MIX-EN	7	0.722	0.74	0.791	0.645
MIX-EN	8	0.712	0.73	0.774	0.63
MIX-EN	9	0.683	0.7	0.741	0.594
MIX-ZH	6	0.928	0.93	0.982	0.919
MIX-ZH	7	0.929	0.93	0.98	0.919
MIX-ZH	8	0.912	0.91	0.974	0.9
MIX-ZH	9	0.911	0.91	0.97	0.9

Figure 8: Complete performance metrics across all training conditions, HuBERT layers, and evaluation settings. This summary includes Accuracy, weighted Precision, ROC-AUC, and F1 scores.