Exploring Machine Learning and Language Models for Multimodal Depression Detection

Javier Si Zhao Hong*, Timothy Zoe Delaya*, Sherwyn Chan Yin Kit*, Pai Chet Ng*, Xiaoxiao Miao†
* Singapore Institute of Technology, Singapore
E-mail: {2302655,2302663,2302669}@sit.singaporetech.edu.sg, paichet.ng@singaporetech.edu.sg
† Duke Kunshan University, China
E-mail: xiaoxiao.miao@dukekunshan.edu.cn

***Abstract*—This paper presents our approach to the first Multimodal Personality-Aware Depression Detection Challenge, focusing on multimodal depression detection using machine learning and deep learning models. We explore and compare the performance of XGBoost, transformer-based architectures, and large language models (LLMs) on audio, video, and text features. Our results highlight the strengths and limitations of each type of model in capturing depression-related signals across modalities, offering insights into effective multimodal representation strategies for mental health prediction.**

I. Introduction

The World Health Organization recently reported that depression affects 3.8% of the global population and 63.6% of these cases remain undiagnosed [1], partly due to the limited availability of healthcare services and, in some cases, financial constraints that prevent many individuals from accessing necessary medical care [2]. Traditional methods, which rely primarily on self-reported questionnaires such as the PHQ-9 [3] and the BDI-II [4], are limited in their ability to capture the dynamic and multifaceted nature of depressive symptoms. These methods are also prone to reporting biases and may fail to detect early or subtle changes in depressive states.

In response to these challenges, the computing community has been instrumental in advancing automatic depression detection by leveraging multimodal data [5], [6], [7], [8], [9], [10], [11], [12], [13]. Most recently, the first Multimodal Personality-Aware Depression Detection (MPDD) Challenge [14] introduced a richly annotated novel dataset that includes audio and visual recordings of participants engaging in a variety of real-world scenarios. The MPDD dataset is annotated using the PHQ-9 scale, Big Five personality traits [15], and detailed demographic information. Compared to existing corpora, the MPDD dataset offers greater contextual diversity and annotation depth, enabling more inclusive and fine-grained modeling.

In terms of methodological approaches, current research increasingly integrates audio, visual, and textual biomarkers through progressively sophisticated computational paradigms. Given that depression detection is a form of fine-grained, subtle emotion recognition, many researchers have drawn inspiration from emotion classification methods. Traditional machine learning approaches typically rely on handcrafted feature extraction pipelines. For example, OpenSMILE-derived acoustic features [16] are often combined with facial expression metrics extracted using computer vision tools, such as Action Units from OpenFace [17] or emotion probabilities from facial emotion recognition models [18]. These features are typically fed into ensemble classifiers, such as XGBoost [19], or kernel-based methods, like support vector machine [20]. Some studies further enhance performance using feature selection or principal component analysis (PCA)-based dimensionality reduction prior to classification [21].

More advanced systems employ hybrid architectures with modality-specific processing pipelines. A common configuration involves using separate convolutional neural networks (CNNs) for visual frames, long short-term memory (LSTM) networks for audio spectrograms, and transformer networks for textual inputs [22]. Fusion strategies vary, ranging from early fusion of low-level features to late fusion of modality-specific predictions [23]. Recent work has also explored cross-modal attention mechanisms using transformers to learn joint representations [24]. Further progress has been achieved through unified architectures that combine self-supervised audio encoders for paralinguistic feature extraction, vision transformers for modeling spatio-temporal facial dynamics, and specialized language models for clinical text analysis [25]. In recent years, large language models (LLMs) have begun to reshape the field. Multimodal emotion recognition systems such as Emotion-LLaMA [26] integrate audio, visual, and textual inputs through emotion-specific encoders. By aligning these features within a shared latent space and applying a modified LLaMA model with instruction tuning, Emotion-LLaMA significantly enhances both emotional recognition and reasoning capabilities.

Among these approaches, XGBoost, transformer-based models, and LLMs have each achieved state-of-the-art results at different stages of research. However, the MPDD Challenge launched in 2025 introduces new complexities that may affect model performance when applying different methods. Inspired by this, the present study systematically evaluates and compares the effectiveness of three representative model classes, XGBoost, transformer-based models, and LLMs, on the MPDD dataset. Following the official challenge protocol, we assess the strengths and limitations of each model across

Xiaoxiao Miao is the corresponding author and this work was conducted while she was at SIT.

979-8-3315-7206-8/25/$31.00

TABLE I
CLASS DISTRIBUTION FOR MPDD DATASET

Task	Label	Elderly		Young	
		#Samples (Ratio)	#Spk	#Samples (Ratio)	#Spk
Binary	Normal	258 (76.6%)	68	135 (51.1%)	45
	Depressed	79 (23.4%)	21	129 (48.9%)	43
	Total	337	89	264	88
Trinary	Normal	138 (40.9%)	37	135 (51.1%)	45
	Mild	120 (35.6%)	31	99 (37.5%)	33
	Severe	79 (23.4%)	21	30 (11.4%)	10
	Total	337	89	264	88
Quinary	Normal	235 (69.7%)	62	-	-
	Mild	68 (20.2%)	18	-	-
	Moderate	23 (6.8%)	6	-	-
	Severe	8 (2.4%)	2	-	-
	Very Severe	3 (0.9%)	1	-	-
	Total	337	89	-	-

TABLE II
FEATURE MODALITIES AND DIMENSIONS IN MPDD DATASET

	Feature Type	**Dimensions**
Audio	MFCC[1]	64
	OpenSMILE[2]	6,373
	Wav2Vec2[3]	512
Visual	DenseNet[4]	1,024 (Elderly) / 1,000 (Young)
	ResNet[5]	1,000
	OpenFace[6]	709
Text	*Raw personality traits for MPDD-Elderly:*	
	Big five: extraversion, agreeableness, openness, neuroticism, conscientiousness	
	Disease category: healthy, other, endocrine, circulatory, neurologica	
	Financial stress: none, mild, moderate, severe/unbearable	
	Family members: number of cohabiting individuals	
	Raw personality traits for MPDD-Young:	
	Big five, Age, Gender, Native place	
	Personalized feature derived from raw personality traits:	
	RoBERTa-large[7]	1,024

modalities, with the goal of identifying their respective potentials for advancing real-world depression recognition systems.

II. THE FIRST MULTIMODAL PERSONALITY-AWARE DEPRESSION DETECTION CHALLENGE

This section provides an overview of the MPDD setup, including the datasets, available audio, visual, and text features, as well as the baseline system, which will serve as the foundation for this study.

A. Datasets

The MPDD dataset comprises two tracks corresponding to distinct age groups, MPDD-Elderly and MPDD-Young, designed to facilitate age-specific depression analysis. Table I provides a comprehensive summary of class distributions across three classification tasks (binary, trinary, and quinary) for both subsets, reporting sample counts, class ratios, and the number of unique speakers (patients)[1].

1) Track 1: MPDD-Elderly: Track 1 focuses on depression detection among elderly participants (average age: 62.8 $\pm$ 11.0). Data were collected through semi-structured interviews conducted in hospital settings. Each participant completed standardized clinical questionnaires, including the PHQ-9 and HAMD-24 scales [27], to assess depression severity. HAMD-24 scores are used to generate labels for the binary and ternary classification tasks, while PHQ-9 scores are used for the quinary classification task.

To enable a more comprehensive participant profile, additional annotations are provided, including Big five personality traits (using a 10-point scale) [15], physical health conditions, financial stress levels, and the number of cohabiting family members, see Table II.

2) Track 2: MPDD-Young: Track 2 targets a younger population (average age: 20.0 $\pm$ 2.2), recruited in non-clinical environments. The data collection protocol consists of a self-introduction, a questionnaire segment, and a scripted reading task, all recorded via video. Participants completed the PHQ-9 questionnaire, which is used to generate depression labels for the binary and ternary classification tasks.

Personality trait annotations in this track differ slightly from those in MPDD-Elderly. In addition to the Big five traits, demographic variables such as age, gender, and place of origin are included, allowing for comparative analysis across different population groups, see Table II.

B. Feature Modalities

Audio features include Mel-frequency cepstral coefficients (MFCCs), low-level acoustic descriptors extracted using OpenSMILE [16], and deep learning-based representations from pre-trained models such as Wav2Vec 2.0 [28]. These features provide a comprehensive view of both handcrafted and learned paralinguistic cues. Visual features comprise deep CNN-based facial embeddings obtained using architectures like DenseNet and ResNet [29], [30], as well as facial behavior analysis (e.g., eye gaze, and head pose) extracted using OpenFace [17]. For the textual modality, both tracks offer RoBERTa-based embeddings [31] derived from raw personality traits descriptions. Each feature type is provided as a fixed-length embedding per 1-second or 5-second hopping window. These variable-length sequences are temporally aligned on a per-subject basis to maintain consistency across modalities.

C. Baseline System

The official baseline system adopts a multimodal deep learning approach, as illustrated on the left of Figure 1. Audio and visual features are first passed through modality-specific encoders, implemented as one-layer LSTM. An optional personalized feature, extracted from a RoBERTa-large model, can be concatenated with the LSTM-processed audio and visual embeddings. The fused representation is then passed

[1]Note that the table only lists the statistics of the MPDD training set. As the time we are writing, the labels of test set are not available. In the following experiments, we split the training set into a 90-10 ratio, using 10% as the development set and reporting the results based on this split.

[1]https://github.com/librosa/librosa
[2]https://github.com/audeering/opensmile
[3]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[4]https://github.com/liuzhuang13/DenseNet
[5]https://huggingface.co/microsoft/resnet-50
[6]http://multicomp.cs.cmu.edu/resources/openface/
[7]https://huggingface.co/FacebookAI/roberta-large

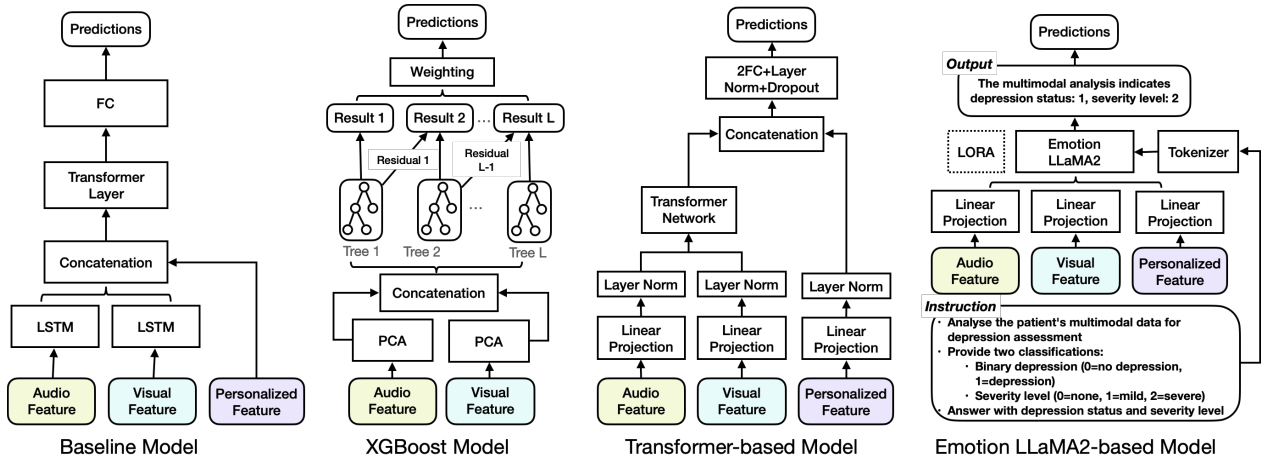

Fig. 1. Various multimodal depression detection models. The leftmost is the baseline model, while the right three are the models investigated in this paper.

through a one-layer transformer that integrates the multimodal information. Finally, the output is fed into fully connected layers to predict the number of classes corresponding to depression severity levels. The system is trained end-to-end using a combination of cross-entropy loss and focal loss [32], and is evaluated using weighted/unweighted F1 scores and overall accuracy.

III. Investigated Systems

This section elaborates on the systems we explored for multimodal depression detection, covering traditional machine learning models, deep learning models, and LLMs.

A. XGBoost-Based Model

We implemented a gradient-boosted decision tree pipeline using XGBoost for multimodal depression classification based on audio and visual features[2], as illustrated in the second panel of Figure 1. Each input sample consists of fixed-length, pre-extracted embeddings derived from pretrained models and represent 1s or 5s window-level summaries of audio and video segments, denoted as $X_a \in \mathbb{R}^{n\times d_a}$ and $X_v \in \mathbb{R}^{n\times d_v}$, respectively, where n is the number of frames for the audio and visual streams, respectively, and d_a, d_v are the corresponding feature dimensions.

To reduce redundancy and improve generalization, we applied Principal Component Analysis (PCA) separately to each modality. For each $X_m \in \{X_a, X_v\}$, we centered the data, computed the covariance matrix C_m, and extracted the top-k eigenvectors V_m^k. Each modality was then projected as $Z_m = (X_m - \text{mean}(X_m))V_m^k \in \mathbb{R}^{n\times k}$, with $k = 50$. The reduced audio and visual features Z_a and Z_v were concatenated to form a fused multimodal embedding $Z = [Z_a \| Z_v] \in \mathbb{R}^{n\times 100}$, which served as input to the XGBoost classifier.

XGBoost is trained for T boosting rounds. At each round t, the model computes the gradients g_t and Hessians h_t of the multi-class log loss with respect to the current predictions $\hat{y}^{(t-1)}$, fits a regression tree f_t to predict the gradients, and updates the predictions as: $\hat{y}^{(t)} \leftarrow \hat{y}^{(t-1)} + f_t(Z)$. The final model aggregates all trees: $\hat{y}(x) = \sum_{t=1}^{T} f_t(x)$.

Besides XGBoost, to address class imbalance, we apply *class weighting* by assigning a higher weight to the minority class. Specifically, the positive class weight is computed as: $w_{\text{pos}} = \frac{N_{\text{neg}}}{N_{\text{pos}}}$, where N_{neg} and N_{pos} are the number of negative and positive samples, respectively. For example, with 90 negatives and 10 positives, $w_{\text{pos}} = 9$. This helps the model better learn from underrepresented classes.

B. Transformer-Based Model

We design a multimodal transformer model to fuse audio, visual, and text features for depression classification.

Each input sample i contains audio $X_a^{(i)} \in \mathbb{R}^{T_a\times d_a}$, visual $X_v^{(i)} \in \mathbb{R}^{T_v\times d_v}$, and text features $x_t^{(i)} \in \mathbb{R}^{d_t}$. These inputs are projected into a shared latent space of dimension d using learned linear layers followed by layer normalization:

$$Z_a^{(i)} = \text{LayerNorm}(X_a^{(i)}W_a + b_a) \in \mathbb{R}^{T_a\times d},$$
$$Z_v^{(i)} = \text{LayerNorm}(X_v^{(i)}W_v + b_v) \in \mathbb{R}^{T_v\times d},$$
$$z_t^{(i)} = \text{LayerNorm}(x_t^{(i)}W_t + b_t) \in \mathbb{R}^{d},$$

where $W_a \in \mathbb{R}^{d_a\times d}$, $W_v \in \mathbb{R}^{d_v\times d}$, and $W_t \in \mathbb{R}^{d_t\times d}$ are learnable projection matrices. Positional encodings $P_a \in \mathbb{R}^{T_a\times d}$ and $P_v \in \mathbb{R}^{T_v\times d}$ are added to preserve temporal order:

$$\hat{Z}_a^{(i)} = Z_a^{(i)} + P_a, \quad \hat{Z}_v^{(i)} = Z_v^{(i)} + P_v.$$

The temporally encoded sequences are passed through modality-specific transformer encoders:

$$H_a^{(i)} = \text{Transformer}(\hat{Z}_a^{(i)}) \in \mathbb{R}^{T_a\times d},$$
$$H_v^{(i)} = \text{Transformer}(\hat{Z}_v^{(i)}) \in \mathbb{R}^{T_v\times d}.$$

To obtain fixed-length representations, we apply learned attention pooling over time:

$$\tilde{x}_a^{(i)} = \sum_{t=1}^{T_a} \alpha_{a,t} H_{a,t}^{(i)} \in \mathbb{R}^d,$$
$$\tilde{x}_v^{(i)} = \sum_{t=1}^{T_v} \alpha_{v,t} H_{v,t}^{(i)} \in \mathbb{R}^d.$$

[2]We attempted to incorporate personalized features but did not observe performance improvements; thus, we excluded them from the XGBoost model.

The pooled audio, pooled visual, and text features are concatenated to form the multimodal representation: $z^{(i)} = \left[\tilde{x}_a^{(i)} \parallel \tilde{x}_v^{(i)} \parallel z_t^{(i)}\right] \in \mathbb{R}^{3d}$. Finally, $z^{(i)}$ is passed through two fully connected layers with ReLU activation, layer normalization, and dropout to predict depression severity. The focal loss function is used to handle class imbalance.

It is well recognized that deep learning models are prone to overfitting when trained on small datasets, as is the case with the MPDD dataset. To address this, we apply *Mixup* data augmentation [33] during training, which has been shown to improve model robustness and performance across various deep learning tasks. The mixup augmentation strategy creates synthetic training examples by linearly interpolating between pairs of samples and their corresponding labels. Specifically, a new example is generated by combining the two input samples and labels using a mixing coefficient.

C. LLM-Based Model

We also explore the use of LLM-based methods for the MPDD task. Specifically, we are inspired by Emotion-LLaMA [26], a model built upon the LLaMA backbone and fine-tuned on large-scale multimodal emotion datasets. Emotion-LLaMA enables emotion reasoning by integrating visual, auditory, and textual cues through structured prompts and cross-modal attention.

Building upon this foundation, we adapt Emotion-LLaMA to the MPDD task by fine-tuning it on our multimodal dataset. The model formulation is expressed as:

$$O = \phi\big(\sigma_{\text{aud}}(X_a^{(i)}), \sigma_{\text{vis}}(X_v^{(i)}), \sigma_{\text{txt}}(x_t^{(i)}), \text{Tokenizer}(\text{Prompt})\big). \tag{1}$$

Here, the input consists of audio features $X_a^{(i)}$, visual features $X_v^{(i)}$, textual features $x_t^{(i)}$, and a task-specific prompt in a multiple-choice question format (as illustrated on the rightmost side of Figure 1).

To integrate features from multiple modalities, we introduce a linear projection mechanism that maps each modality into a shared embedding space. This is achieved via trainable linear projection functions: σ_{aud} for audio, σ_{vis} for visual, and σ_{txt} for text. The final output O is a formatted text response, also shown in the rightmost bottom of Figure 1.

The fine-tuning process involves two stages. In Stage 1, the LLaMA backbone is frozen, and only the projection layers and classification heads are trained. In Stage 2, LoRA-based fine-tuning is applied, using a dual learning rate setup to update both the LoRA parameters and the projection layers.

IV. Experiments

Experiments begin with the 5-second MPDD-Elderly dataset, where we conduct a comprehensive ablation study to identify the optimal configuration for each system. Once the best settings are determined, these configurations are applied to the remaining scenarios. For all scenarios, we use a 90-10 train-validation split based on patient IDs, ensuring subject independence and preventing information leakage. We evaluated each configuration using two primary metrics: weighted (W_{F1}) and unweighted F1 scores (U_{F1}).

[1]The parameters of the baseline system are calculated using wav2vec and OpenFace features as the input features.

TABLE III
Ablation study on each systems for 5s MPDD Elderly binary dev set (90(cross-validation)/10).

Methods	W_{F1}	U_{F1}
XGBoost		
Raw Feature	65.76	42.62
Raw Feature + *class weighting*	74.80	65.04
PCA Feature + *class weighting*	94.29	91.90
Transformer		
2 transformer layers	80.44	69.64
2 transformer layers + *mixup*	84.00	76.22
2 transformer layers +*mixup* + *cross-validation*	87.08	83.19
LLM		
Llama2	60.96	44.57
EmotionLlama2 + 1step	31.77	33.33
EmotionLlama2 + 2steps	70.59	52.55

A. Experiment Settings

1) XGBoost Setting: The XGBoost configuration was carefully designed to balance interpretability, computational efficiency, and robust performance on limited multimodal data. Each audio and visual modality was first reduced to 50 dimensions using PCA, resulting in a fused 100-dimensional feature vector. For configurations without PCA, original modality features were concatenated directly. The model used a shallow tree depth with a maximum depth of 3, a learning rate selected from $\{0.01, 0.05\}$, and both subsample and colsample-by-tree ratios set to 0.8 to introduce randomness and reduce overfitting. Training employed up to 500 boosting rounds with early stopping after 25 rounds without improvement, using multi-class log loss as the evaluation metric and `multi:softprob` as the objective function. Hyperparameters were tuned per modality pair and classification task (binary, ternary, quinary) using early stopping on a speaker-level validation split to avoid overfitting and data leakage. Multiple audio-visual fusion pairs were explored (e.g., MFCC + OpenFace, OpenSMILE + ResNet), with MFCC + OpenFace achieving the best results. Consistency was maintained by applying the same patient-level ID split and evaluation methodology across all experiments. Notably, personality-aware features and probabilistic model ensembling were excluded to enable a focused evaluation of core modality fusion performance under classical machine learning settings.

2) Transformer setting: The transformer configuration prioritizes efficiency and regularization to avoid overfitting on the relatively small dataset. The model uses a reduced dimensionality of 128, with a shallow 2-layer transformer architecture and 4 attention heads, which together provide sufficient representational capacity while maintaining computational efficiency. A dropout rate of 0.5 is applied throughout the network for strong regularization. The pre-classifier network reduces the concatenated multimodal embedding from 512 to 256 dimensions, refining the joint feature representation for

TABLE IV
WEIGHTED F1 AND UNWEIGHTED F1 (%) ↑ RESULTS ON MPDD-ELDERLY DEV SET.

Method	PF	1s						5s					
		Binary		Ternary		Quinary		Binary		Ternary		Quinary	
		W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}
Baseline	✗	82.60	70.89	54.35	49.14	63.85	44.00	77.90	66.15	50.88	47.59	73.49	56.83
Baseline	✓	85.71	79.13	56.48	55.64	66.26	46.66	81.75	72.37	58.22	59.37	75.62	58.40
XGBoost	✗	90.67	85.83	55.23	53.02	55.60	21.43	**94.29**	**91.90**	61.02	**62.51**	54.62	21.05
Transformer	✓	**93.44**	**88.21**	**74.95**	**80.00**	**82.21**	**46.77**	85.27	71.55	**65.52**	61.31	67.96	**67.62**
LLM	✓	70.59	52.55	61.17	53.02	77.89	30.60	67.43	45.60	45.66	37.44	**77.89**	30.60

TABLE V
WEIGHTED F1 AND UNWEIGHTED F1 (%) ↑ RESULTS ON MPDD-YOUNG DEV SET WITH MODEL SIZE FOR THE BINARY TASK.

Method	PF	#Params(M)	1s				5s			
			Binary		Ternary		Binary		Ternary	
			W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}	W_{F1}	U_{F1}
Baseline	✗	1.89[1]	55.23	55.23	47.95	43.72	60.02	60.02	42.82	39.38
Baseline	✓	2.15	59.96	59.96	51.86	51.62	62.11	62.11	48.18	41.31
XGBoost	✓	0.002	81.53	81.38	66.67	48.89	74.07	74.07	62.19	45.60
Transformer	✓	1.06	**95.83**	**95.83**	**75.60**	**71.36**	**81.48**	**81.48**	**78.51**	**59.16**
LLM	✓	6,843	60.86	61.65	45.65	34.06	64.07	64.96	39.49	28.82

downstream classification. For training, we use a learning rate of 5e-5, batch size of 2, and train for up to 100 epochs, with early stopping triggered if no improvement is observed for 20 consecutive epochs. We also include warmup training for the first 10 epochs, gradient clipping at 1.0, and apply weight decay of 1e-4 to further aid generalization. Multiple audio-visual fusion pairs were explored, and wav2vec2, DenseNet, with personalized features were selected.

For the mixup augmentation strategy, the mixing coefficient is sampled from a Beta distribution with parameters 0.2 and 0.2. Applied consistently across all modalities with a 50% probability, mixup helps the model learn smoother decision boundaries and improves generalization on limited data. The cross-validation experiment employs 10-fold validation, where each fold naturally provides the 90-10 split used in strategies without cross-validation.

3) LLM setting: The Depression-LLaMA implementation is based on the Emotion-LLaMA pre-trained foundation, utilizing the LLaMA-2-7B[3] as the underlying large language model. The model architecture utilize all features, including three audio features, three visual feature and one text features listed in table II Each of them will be mapped to 4,096 dimensional features by the linear projection. The training process is conducted in two stages. In Stage 1, the backbone is frozen and trained for 5 epochs with a learning rate of 5×10^{-5}. In Stage 2, LoRA fine-tuning is applied for 3 epochs with a learning rate of 1×10^{-5}.

B. Results

1) Ablation Study for each system: Table III presents the weighted F1 (W_{F1}) and unweighted F1 (U_{F1}) scores for various systems evaluated on the 5-second MPDD Elderly binary development set using a 90-10 split (with cross-validation where specified). The results compare baseline methods, traditional machine learning approaches, transformer models, and LLM baselines, with different training strategies.

XGBoost: Starting with raw features, XGBoost achieves moderate performance. Incorporating class weighting leads to a notable improvement, and applying PCA alongside class weighting further enhances results, demonstrating the effectiveness of dimensionality reduction and handling class imbalance.

Transformer: The baseline transformer outperforms XGBoost without PCA, showing strong capability in modeling the data. Adding mixup augmentation further improves the model's generalization, and combining mixup with cross-validation yields the best and most robust performance, highlighting the benefits of data augmentation and rigorous evaluation.

LLM Methods: The baseline LLaMA2 model performs relatively poorly compared to other methods. Initial fine-tuning of EmotionLLaMA2 results in a performance drop, likely due to adaptation challenges, but subsequent fine-tuning improves outcomes considerably. Despite this, LLM-based methods still lag behind the transformer and XGBoost models on this task.

2) Overall Results: Table IV and Table V present a comprehensive comparison of all explored methods on the MPDD-Elderly dataset. On the MPDD-Elderly development set, XGBoost achieves the highest weighted and unweighted F1 scores for the 5-second binary classification task, demonstrating strong performance on longer audio segments with effective feature engineering. The Transformer model outperforms XGBoost on shorter 1-second segments and more complex classification tasks (ternary and quinary), indicating its strength in capturing fine-grained information. Personalized features improve baseline models but the LLM-based approach shows comparatively lower performance across most tasks.

For the MPDD-Young development set, the Transformer consistently delivers the best results across all classification tasks and time windows, with weighted F1 scores reaching as

[3]https://huggingface.co/meta-llama/Llama-2-7b

high as 95.83% on 1-second binary classification. XGBoost performs well but lags behind Transformer models, especially on ternary tasks. Baseline models benefit from personalized features but remain less competitive, while LLM-based methods perform the weakest.

Overall, the Transformer model (1.06M parameters) is the most effective for multimodal depression detection, especially for younger speakers and shorter audio windows. Despite having significantly more parameters, LLM (6,843M) underperforms, especially in ternary classification, suggesting that larger models don't always guarantee better performance. XGBoost, with only 0.002M parameters, achieves strong results in binary classification, highlighting the effectiveness of simpler models for specific tasks. The Baseline with a personalized feature slightly improves on the baseline without it but remains less effective than more complex models, even though it has more parameters (2.15M) than the Transformer.

Acknowledgment This research is supported by the Ministry of Education, Singapore, under its Academic Research Tier 1 (Grant number: GMS 956) and the Academy of Medical Sciences, under its Networking Grant (NGR1\1678).

REFERENCES

[1] World Health Organization, "Depressive disorder (depression)," https://www.who.int/news-room/fact-sheets/detail/depression, 2023, accessed: 7 July 2025.

[2] A. Faisal-Cury, C. Ziebold, D. M. O. Rodrigues, and A. Matijasevich, "Depression underdiagnosis: Prevalence and associated factors. a population-based study," *Journal of Psychiatric Research*, vol. 151, pp. 157–165, July 2022, epub 2022 Apr 23.

[3] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, September 2001, pMID: 11556941.

[4] A. T. Beck, R. A. Steer, and G. K. Brown, *Beck Depression Inventory—Second Edition Manual*. San Antonio, TX: The Psychological Corporation, 1996, © 1996, 1987 by Aaron T. Beck, Robert A. Steer, Gregory K. Brown.

[5] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2017.

[6] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 12 226–12 234.

[7] 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.

[8] 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, 2022.

[9] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li *et al.*, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, no. 1, p. 178, 2022.

[10] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.

[11] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[12] D. N. Klein, R. Kotov, and S. J. Bufferd, "Personality and depression: explanatory models and review of the evidence," *Annual Review of Clinical Psychology*, vol. 7, no. 1, pp. 269–295, 2011.

[13] M.-T. Lo, D. A. Hinds, J. Y. Tung, C. Franz, C.-C. Fan, Y. Wang, O. B. Smeland, A. Schork, D. Holland, K. Kauppi *et al.*, "Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders," *Nature Genetics*, vol. 49, no. 1, pp. 152–156, 2017.

[14] C. Fu, Z. Fu, Q. Zhang, X. Kuang, J. Dong, K. Su, Y. Su, W. Shi, J. Yao, Y. Zhao, S. Zhao, J. Wang, S. Song, C. Liu, Y. Yoshikawa, B. Schuller, and H. Ishiguro, "The first mpdd challenge: Multimodal personality-aware depression detection," arXiv preprint arXiv:2505.10034, 2025, accepted at the MPDD Challenge, ACMMM 2025 Grand Challenge. [Online]. Available: https://arxiv.org/abs/2505.10034

[15] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[17] T. Baltrusaitis *et al.*, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE FG*, 2018, pp. 59–66.

[18] J. Shenk, "Python fer (facial expression recognition)," https://github.com/justinshenk/fer, 2019.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[20] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[21] J. Gideon *et al.*, "Detecting deprefssion on video logs using audiovisual features," *Humanities & Social Sciences Communications*, vol. 10, no. 1, pp. 1–12, 2023.

[22] Y. Zhang *et al.*, "First transformer-based depression detection using multi-head attention," *Sensors*, vol. 21, no. 14, p. 4764, 2021.

[23] J. Smith *et al.*, "Late fusion strategies for multimodal depression classification," *Psychiatry AI*, vol. 12, pp. 45–60, 2023.

[24] L. Chen *et al.*, "Cross-attention multimodal fusion using macbert for depression detection," *arXiv:2407.12825*, 2024.

[25] S. Ji *et al.*, "Mentalbert: A clinical language model for mental health assessment," *Natural Language Processing Journal*, vol. 1, p. 100003, 2022.

[26] Z. Cheng, Z.-Q. Cheng, J.-Y. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 110 805–110 853. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/c7f43ada17acc234f568dc66da527418-Paper-Conference.pdf

[27] M. Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*. Springer, 1986, pp. 143–152.

[28] 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.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2018.