I.J. Image, Graphics and Signal Processing, **2026, 2, 124-136**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.02.08

A Robust Hybrid Deep Learning Model for Multiclass Depression Classification from Speech Audio

Neny Sulistianingsih*
Computer Science, Bumigora University, Mataram, Indonesia
Email: neny.sulistianingsih@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0000-0003-0548-5038
*Corresponding Author

Galih Hendro Martono
Computer Science, Bumigora University, Mataram, Indonesia
Email: galih.hendro@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0000-0002-0697-010X

Received: 17 June, 2025; Revised: 15 November, 2025; Accepted: 02 February, 2026; Published: 08 April, 2026

Abstract: Depression remains one of the most prevalent and underdiagnosed mental health disorders globally, necessitating scalable, objective, and non-invasive diagnostic tools. Speech, as a rich biomarker of emotional and psychological states, offers a promising avenue for automated depression detection. This study proposes a robust hybrid deep learning framework that integrates Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), Bidirectional Long Short-Term Memory (BiLSTM), and Transformer architectures to classify depression severity into three levels: normal, mild, and severe. Using a curated multimodal dataset comprising 400 labeled audio recordings, we extract comprehensive acoustic features, including MFCC, Chroma, Spectrogram, Contrast, and Tonnetz representations. Models are evaluated using precision, recall, F1-score, and accuracy. Experimental results show that the proposed hybrid models outperform traditional architectures, achieving up to 99% accuracy and strong generalization across all classes. This study demonstrates the potential of attention-enhanced hybrid architectures in mental health assessment and provides a foundation for future deployment in clinical and real-world settings. Future work includes multimodal fusion with EEG data and the implementation of explainable AI for clinical interpretability.

Index Terms: Depression Detection, Speech Emotion Recognition, Hybrid Deep Learning, CNN, Transformer, GRU, BiLSTM, Mental Health Assessment

1. Introduction

Mental health disorders, particularly depression, have become a significant global health concern, affecting more than 300 million people worldwide [1]. The increasing prevalence and underdiagnosis of depression have prompted the need for scalable, objective, and non-invasive detection methods. Among various modalities, speech has emerged as a valuable biomarker due to its strong correlation with emotional and psychological states [2, 3]. The acoustic features in speech can reflect cognitive and affective disturbances, making it a promising avenue for automated depression detection systems.

Over the past decade, numerous studies have employed machine learning (ML) and deep learning (DL) techniques for depression classification using speech data. Early approaches utilized traditional classifiers such as Support Vector Machines (SVM) and Random Forests (RF) with handcrafted features like MFCCs, pitch, and jitter [4, 5]. More recently, deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated improved performance in learning discriminative spectral and contextual patterns from speech representations directly from raw or spectrogram-transformed audio [6–9]. For example, [7] proposed a Spectro-CNN hybrid model that achieved promising results on DAIC-WOZ and MODMA datasets.

In parallel, hybrid models that integrate multiple architectures—such as CNNs with LSTMs or GRUs—have also gained traction due to their ability to capture spatial and sequential dependencies [10, 11]. Moreover, attention mechanisms and Transformer-based models have enhanced depression recognition by emphasizing salient temporal regions in speech signals [12]. Although multimodal approaches incorporating textual, visual, or neurophysiological

signals have been investigated in prior literature [13], the present study focuses exclusively on speech-based modeling, with other modalities considered beyond the scope of the current experimental evaluation.

Beyond audio, multimodal approaches incorporating text, facial expression, and EEG data have demonstrated that fusing complementary cues improves classification performance.

Despite these advancements, several research gaps remain. First, while many models focus on binary classification (depressed vs. non-depressed), fewer studies have explored multiclass depression severity detection, which is more clinically relevant [14]. Second, most audio-based models rely solely on CNN or RNN, with limited integration of Transformer architectures for contextual feature interactions. Third, there is a lack of hybrid systems that combine CNNs, RNNs, and Transformer blocks in a unified architecture, especially in non-English, low-resource speech datasets. Finally, the use of EEG as a complementary modality remains underutilized despite its potential for revealing neurophysiological correlates of depression [15].

Rather than proposing a fundamentally new architecture, the study aims to benchmark and analyze the effectiveness of different hybrid configurations that integrate CNN, LSTM, GRU, and Transformer components for multiclass depression severity classification from speech audio. Rather than introducing a fundamentally new architecture, the study aims to analyze and benchmark the effectiveness of different hybrid combinations in a low-resource audio setting. The models are evaluated on a curated dataset of labeled speech recordings categorized into three depression severity levels: normal, mild, and severe. The proposed pipeline includes standardized preprocessing, acoustic feature extraction (MFCC, Chroma, and Tonnetz), and comprehensive evaluation using accuracy, precision, recall, and F1-score.

The main contributions of this work are summarized as follows:

1) We benchmark multiple hybrid and non-hybrid architectures combining CNN, GRU, BiLSTM, and Transformer layers, enabling the extraction of both local and global temporal features from emotional speech.
2) We address the clinical need for severity-aware classification by formulating the task as a multiclass depression classification problem.
3) We benchmark our approach on a unique multimodal dataset with three class labels and demonstrate performance improvements in weighted F1-score and generalization over conventional architectures.
4) We provide an extensive comparison with recent state-of-the-art models, highlighting the effectiveness and scalability of our method.

The remainder of this paper is organized as follows: Section II reviews related work in speech-based depression detection. Section III presents the proposed hybrid model and evaluation pipeline. Section IV describes the experimental results, discussion, implications and limitations. Finally, Section V concludes the study and outlines future research directions.

2. Literature Review

Recent years have witnessed a surge in research utilizing speech and multimodal data for automated depression detection, motivated by the growing accessibility of deep learning (DL) techniques and the urgent demand for scalable mental health screening solutions. [16] developed an Attention-Based Audio Fusion Network (ABAFNet) combining LSTM and attention-based late fusion for depression detection using the CNRAC and CS-NRAC datasets. Their model effectively highlighted the importance of MFCC features in fusion strategies, significantly outperforming traditional baselines. Similarly, [7] proposed a CNN-RNN hybrid using spectrograms and MFCCs, reporting F1-scores of approximately 91% across multiple benchmark datasets (e.g., MODMA, RAVDESS), further affirming the value of deep hybrid architectures in capturing affective cues.

Multimodal learning has also become a dominant approach in recent work. Study by [12] introduced a complex teacher-student framework integrating BiLSTM, Wav2Vec2, and Llama3-8B, achieving an F1-score of 99.1% on the DAIC-WOZ dataset. [3] further demonstrated the utility of additive cross-modal attention in enhancing classification performance across the DAIC-WOZ and EATD-Corpus, reaching 82% F1-score.

CNN-GRU and CNN-BiLSTM models have also proven to be robust in single-modality settings. [6] integrated GAT-based GNNs and attention mechanisms, showing that emotional self-attention significantly improves generalization across datasets. Menawhile [5] explored unsupervised feature learning with CNN autoencoders and achieved notable gains in real-time detection scenarios.

In addition, Transformer-based and lightweight fusion architectures have gained prominence. [17] implemented a hybrid model combining MLP-Mixer and Transformer with cross-attention, obtaining an F1-score up to 0.81. [18] used Conv1D and Transformer layers with attention fusion for scalable screening using audio-visual cues, outperforming existing models like DepAudioNet.

EEG-based approaches have also emerged. [15] introduced an interpretable hybrid EEG model using ALO-MARL and SHAP, achieving a high F1-score of 91.82%, while [19] integrated EEG with vision and speech using Transformer networks to reach F1-scores near 97%.

Several reviews by [2, 20] emphasized that attention pooling, hybrid feature extraction, and personalization are essential for improving speech-based depression recognition. Meanwhile, traditional RNN and CNN models remain competitive when fused with modern components like attention or domain-aligned embeddings, as demonstrated by [14, 21].

Despite the wide variety of models, few studies have focused on multiclass classification that reflects real-world clinical depression severity levels. Moreover, limited works integrate CNN, RNN, and Transformer components in a unified architecture exclusively for speech-based depression analysis. This research aims to bridge this gap by proposing a robust hybrid deep learning framework that leverages the strengths of CNNs, GRUs, BiLSTMs, and Transformers for multiclass depression detection using emotional speech recordings—a state-of-the-art current study as shown in Table 1.

Table 1. Comparison with recent dtate-of-the-art studies in depression detection.

Study	Dataset Type	Multimodal	Audio-Only	EEG	Deep Learning	Hybrid CNN-RNN	Transformer	Attention	Multiclass	Label Type
[16]	Speech Only	Yes	Yes	No	Yes	Yes	No	Yes	No	Binary
[12]	DAIC-WOZ	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Binary
[3]	DAIC-WOZ/EATD	Yes	Yes	No	Yes	Yes	No	Yes	No	Binary
[7]	DAIC-WOZ/MODMA	No	Yes	No	Yes	Yes	No	Yes	No	Binary
[6]	Various Audio	No	Yes	No	Yes	Yes	No	Yes	No	Binary
[4]	Survey Tabular	No	No	No	Yes	No	No	No	No	Binary
[9]	DAIC-WOZ	No	Yes	No	Yes	Yes	Yes	Yes	No	Binary
[22]	Video + Audio + rPPG	Yes	Yes	No	Yes	No	Yes	Yes	No	Regression
[23]	WIBD	Yes	Yes	No	No	No	No	No	No	Regression
[10]	DAIC-WOZ	No	Yes	No	Yes	Yes	No	No	No	Binary
Current Study	Audio	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Multiclass

3. Research Methodology

This study proposes a robust pipeline for depression classification based on emotional speech features. The methodology consists of several systematic stages: data acquisition, feature extraction, model construction using deep learning and hybrid architectures, and evaluation using standard performance metrics. Fig. 1 illustrates the overall architecture of the proposed method.

3.1. Data Acquisition

The datasets utilized in this study are sourced from the KaggleHub repository, organized under the Multimodal Dataset for Depression Analysis project [24]. These datasets provide diverse modalities—namely, audio and electroencephalogram (EEG)— to enable robust modeling of psychological states and to support both classification and regression tasks in mental health research. While relatively modest in size, these datasets have been successfully used in prior studies focused on emotion recognition and affective computing, demonstrating their value in early-stage prototyping and model benchmarking. Furthermore, several recent studies have shown that even with limited data, deep learning models—when coupled with appropriate regularization and data augmentation strategies—can achieve strong generalization, particularly in low-resource mental health applications [25, 26].

The directory includes three main datasets, each targeting a unique aspect of emotional and cognitive processing:

1) Multimodal Emotional Speech and Song Dataset

Originating from the Department of Psychology at Ryerson University, this dataset contains 288 audio clips of 3-second duration each, recorded by 24 professional actors. The recordings cover eight distinct emotional states: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised. The dataset is widely used for emotion recognition, mental health assessment, and multimodal affective computing [27].

2) Audio Dataset for Depression Analysis

This curated collection consists of 400 speech recordings in .wav format. The tapes capture diverse emotional tones and vocal patterns indicative of psychological states. Each sample is labeled into one of three classes: Not Depressed, Mildly Depressed, and Severely Depressed. This dataset is well-suited for supervised machine learning, deep learning classification, and speech-based depression detection systems [28].

3) Depression EEG Dataset

Comprising task-related EEG signals from 30 diagnosed depressed individuals, this dataset is intended to support neuroscience and bio-signal processing studies. The recordings capture task-state brain activity focusing on negative wave patterns in response to false associations. Additionally, it enables comparative analysis between depressed patients and healthy or OCD-affected individuals. It helps explore cortical response mechanisms related to affective disorders [29].

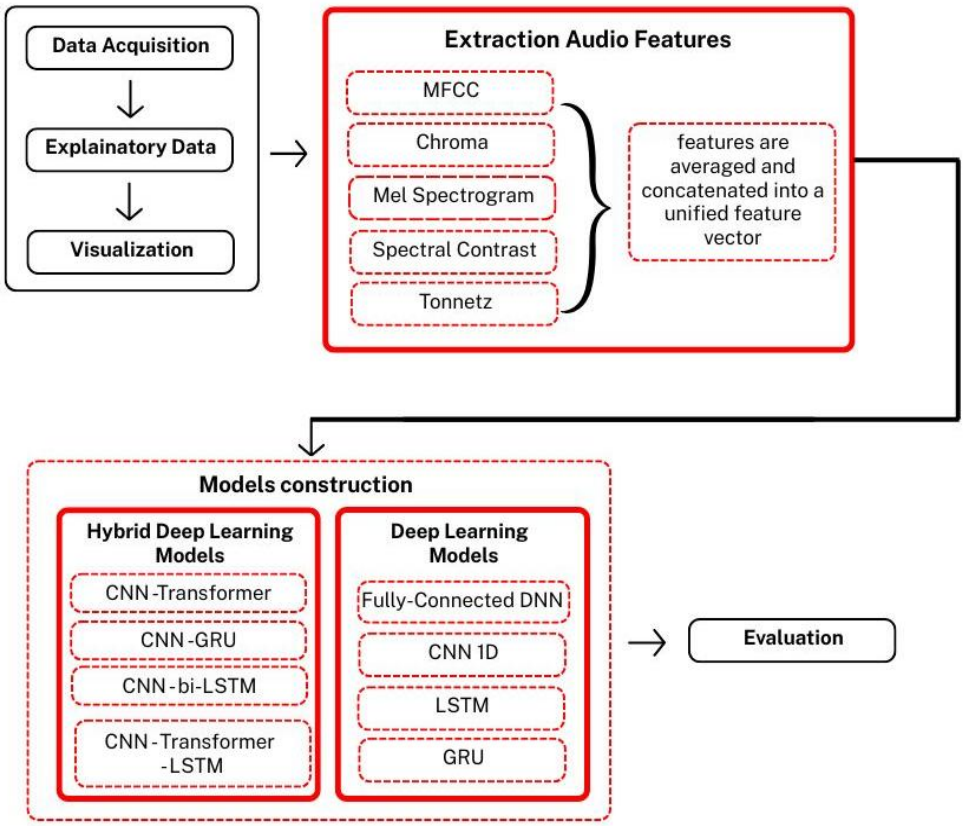

Fig. 1. Architecture of the proposed method.

All datasets are pre-organized into directory structures based on class labels, ensuring compatibility with deep learning pipelines. After downloading the data using the KaggleHub Python API, the directory structure was verified through recursive file tree parsing to ensure data completeness. For the audio-based depression dataset, samples were organized into three class folders corresponding to Normal, Depression Stage 1, and Depression Stage 2.

For model evaluation, the audio dataset was partitioned into training and testing sets using an 80/20 split at the sample level. It is important to note that the dataset does not provide explicit speaker identity metadata. Consequently, speaker-independent partitioning could not be strictly enforced, and recordings from the same speaker may potentially appear in both the training and testing sets. This limitation is acknowledged as a potential source of optimistic performance estimation and is discussed further in the Discussion section. As such, the reported results should be interpreted as an exploratory evaluation rather than a definitive assessment of cross-speaker generalization.

Exploratory analysis was subsequently conducted by visualizing representative samples using time–frequency spectrograms, enabling qualitative inspection of acoustic differences across depression severity levels. Despite the dataset size and structural limitations, the diversity of emotional expressions captured in the speech signals provides a useful foundation for investigating the behavior of hybrid deep learning architectures in low-resource depression detection settings.

3.2. Extraction of Audio Features

Audio-based depression detection relies heavily on the quality and relevance of the acoustic features extracted from speech recordings. In this study, we performed comprehensive feature extraction from .wav audio files using the Librosa library in Python, a widely adopted toolkit for music and audio analysis [30].

Five distinct types of features were extracted to represent various dimensions of acoustic and perceptual information:

1) Mel-Frequency Cepstral Coefficients (MFCCs) are among the most informative features for modeling the human vocal tract and are considered the de facto standard in speech and affective computing tasks. They effectively capture the short-term power spectrum of a sound, closely aligning with human auditory perception.
2) Chroma Features represent energy distribution across the 12 distinct pitch classes (semitones of the musical octave), which are particularly useful for capturing tonal characteristics and intonation, elements known to vary in depressed speech [31].
3) Mel Spectrogram provides a time-frequency representation of sound, offering insight into the spectral structure of speech over time. It is commonly used as an input for deep learning models due to its visual similarity to natural images [25].
4) Spectral Contrast captures the difference in amplitude between peaks and valleys in the sound spectrum. Depressed individuals often exhibit flatter prosody and reduced spectral variation, making spectral Contrast a valuable marker [26].
5) Tonnetz (Tonal Centroid Features) was initially developed in music theory. Tonnetz features capture harmonic relations between pitches and have helped model emotional valence in speech [32].

Each extracted feature matrix was temporally averaged to produce fixed-length feature vectors. This design choice was adopted to standardize input dimensionality across samples of varying duration and to ensure stable training in a low-resource setting, where fully sequence-preserving representations may increase model variance and overfitting risk. Although recurrent and Transformer-based architectures are inherently designed to model temporal dependencies, in this study they were employed to learn higher-order relationships among aggregated acoustic descriptors, rather than frame-level temporal dynamics. This formulation enables a controlled comparison between different architectures under uniform input conditions, while prioritizing robustness and reproducibility given the dataset size.

The temporally aggregated feature vectors were then concatenated into a single composite representation, allowing uniform input across different deep learning models regardless of the original length of the audio clips. To aid interpretability, representative spectrograms were visualized for each class label (Normal, Stage 1 Depression, Stage 2 Depression), demonstrating distinctive patterns in time–frequency structure and spectral energy. All subfolders were programmatically accessed and validated to ensure completeness before feature extraction. Exploratory analysis was subsequently conducted by visualizing representative samples, enabling qualitative insights into emotional differences encoded in the speech signals. Although limited in size, the diversity of emotional expressions captured in the dataset provides a useful foundation for investigating the behavior of hybrid deep learning models in low-resource mental health assessment scenarios.

3.3. Models Construction

Consistent with the feature extraction strategy described in Section 3.2, all deep learning models in this study operate on temporally aggregated acoustic representations rather than frame-level sequences. To classify depression stages from audio recordings, we constructed and evaluated multiple deep learning architectures, ranging from simple dense neural networks to more sophisticated hybrid models. Each architecture was carefully designed to leverage different feature representations and contextual relationships inherent in speech signals, rather than explicit frame-wise temporal dynamics.

First, we implemented a fully connected deep neural network (DNN) as a baseline model to assess the effectiveness of high-level aggregated audio features. This model consists of stacked dense layers, with batch normalization and dropout, to reduce overfitting and enhance generalization [33].

We utilized a 1D Convolutional Neural Network (CNN-1D) to capture localized patterns within the aggregated acoustic feature space. CNNs are well-known for extracting salient patterns in sequential data by applying convolutional filters and pooling layers, which have been effective in speech-processing tasks [34].

Next, we introduced Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. These architectures are designed to model temporal dependencies and sequential dynamics in time-series data such as speech. While these architectures are traditionally used for modeling temporal dependencies in sequential data, in this study they are employed to model relational and contextual interactions among aggregated acoustic feature dimensions. LSTMs provide higher representational capacity for capturing complex feature relationships [35], whereas GRUs provide a computationally lighter alternative with comparable performance [36].

We implemented a Hybrid Transformer-based model to enhance further modeling capacity, which integrates self-attention mechanisms to capture global contextual relationships within audio features. In this formulation, the Transformer block is used to capture global contextual relationships among aggregated audio features, rather than explicit frame-level temporal dependencies. This allows the model to emphasize informative feature interactions through multi-head attention, extending beyond the capabilities of conventional RNN-based architectures [37]. To further explore synergistic architectural benefits, we constructed hybrid models integrating convolutional layers for feature extraction with sequential modeling components (RNN/Transformer) to enhance context-aware feature interaction learning.

The CNN-GRU model combines a convolutional front-end with a GRU backend. The Conv1D layer extracts local features from audio sequences, which are then passed to a GRU layer to capture medium-range dependencies in the signal. This architecture offers a trade-off between expressiveness and computational efficiency [38].

The CNN-BiLSTM model introduces bidirectional LSTM layers after the convolutional layer. Unlike standard LSTMs that process sequences in one direction, BiLSTMs process information in both forward and backward directions, allowing the model to capture context from past and future time steps [39]. This is particularly advantageous in speech analysis, where cues relevant to depression may appear throughout the audio segment.

Lastly, the CNN-Transformer-LSTM model incorporates a three-stage design: a CNN block for localized feature extraction, a Transformer block for long-range dependency modeling, and an LSTM to refine sequential representations. The Transformer module's multi-head attention lets the model focus on multiple audio parts simultaneously, effectively learning global dependencies across frames [37].

Each model is compiled with the Adam optimizer and categorical cross-entropy loss, which is suitable for multiclass classification. To maintain consistency, these hybrid models were evaluated with the same dataset and metrics as the standalone models. Their parameter counts and architectural structures are summarized in Table 2.

Table 2. Summarize the model's parameter and its architectural structures in the current study.

Model Name	Architecture Summary	Parameters	Complexity	Remarks
Fully-Connected DNN	Dense(256)-BN-Dropout-Dense(128)-Dropout-Dense(3)	77,059	Low	Tabular feature classifier
CNN-1D	Conv1D(64)-MP1D-Flat-Dense(128)-Dropout-Dense(3)	672,515	Medium	Local temporal feature extractor
LSTM	LSTM(64)-LSTM(32)-Dense(128)-Dropout-Dense(3)	33,923	High	Long-term sequence modeling
GRU	GRU(64)-GRU(32)-Dense(128)-Dropout-Dense(3)	26,883	High	Gated RNN for efficient sequence modeling
CNN-Transformer	Dense(256)-ExpandDims-TransformerBlock-Flat-Dense(128)-Dropout(3)	1,194,883	Very High	Global sequence attention
CNN-GRU	Conv1D(64)-MP1D-GRU(64)-Flat-Dense(64)-Dropout-Dense(3)	365,443	Medium-High	Efficient hybrid model combining spatial & temporal info
CNN-BiLSTM	Conv1D(64)-MP1D-BiLSTM(64)-Flat-Dense(64)-Dropout-Dense(3)	746,499	High	Context-aware bidirectional memory
CNN-Transformer-LSTM	Conv1D(64)-MP1D-LN-MHA-LSTM(64)-Flat-Dense(64)-Dropout-Dense(3)	440,003	Very High	Strong global attention + sequence learning

3.4. Evaluation

Several evaluation metrics were employed to assess the proposed deep learning models' performance objectively. These metrics are widely used in multiclass classification tasks [40], especially in the domain of speech and affective computing [41, 42]. The metrics include Accuracy, Precision, Recall, and F1-Score, each capturing a different aspect of classification performance.

Let TP, FP, FN, and TN denote the number of true positives, false positives, false negatives, and true negatives for a particular class. The following definitions are extended for a multiclass scenario using a macro- or weighted-averaging approach over all classes.

Accuracy measures the overall proportion of correct predictions across all classes, so

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In the multiclass case:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} TP_i \tag{2}$$

$$N = \sum_{i=1}^{C} \left(TP_i + FP_i + FN_i \right) \tag{3}$$

Precision quantifies the number of true positive predictions out of all predicted positive instances, so

$$\Pr. = \frac{TP}{TP + FP} \tag{4}$$

Meanwhile, for multiclass classification, a weighted precision is computed as (5).

$$\mathrm{Pr}_{weighted} = \sum_{c=1}^{C} w_c \times \frac{TP_c}{TP_c + FP_c} \tag{5}$$

where ω_c is the class-wise weight, typically proportional to the support of class c, and C is the total number of classes.

Recall (also known as sensitivity) captures how well the model identifies actual positive instances so

$$\mathrm{Re.} = \frac{TP}{TP + TN} \tag{6}$$

In a weighted multiclass setting:

$$\mathrm{Re}_{weighted} = \sum_{c=1}^{C} w_c \times \frac{TP_c}{TP_c \times FN_c} \tag{7}$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure even when class distribution is imbalanced, so

$$FIscore = 2 \times \frac{\mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{8}$$

The weighted F1-score used in this study accounts for class imbalance:

$$F1score_{weighted} = \sum_{c=1}^{C} w_c \times F1_c \tag{9}$$

This comprehensive metric provides a single performance score that balances false positives and false negatives.

These metrics ensure that model correctness and sensitivity to minority classes are evaluated fairly, especially given the class imbalance typical in psychological health datasets [43].

All metrics were calculated using the sklearn—metrics library and evaluated on the held-out test set (20% split). Confusion matrices were also used for per-class analysis to observe the tendency of misclassifications. This combination of metrics ensures robust evaluation for the proposed models, as demonstrated in recent research on depression detection using multimodal and speech-based approaches [44].

4. Results and Discussion

4.1. Result

All models were trained using identical optimization settings, including the same optimizer, loss function, batch size, and number of epochs, without extensive per-model hyperparameter tuning, to ensure a fair architectural comparison. This section presents a comparative analysis of all models developed in this study for depression-level classification based on audio data. The performance of each model was evaluated using four commonly used classification metrics: Precision, Recall, F1-Score, and Accuracy. The evaluation was conducted per class label—Normal, Depression Stage 1, and Depression Stage 2—as shown in Table 3. Meanwhile, graphic loss and accuracy for CNN-GRU are shown in Fig. 2. It should be noted that model evaluation was conducted using a sample-level train–test split due to the absence of explicit speaker identity metadata, and therefore the reported performance does not reflect strict speaker-independent generalization.

As observed, hybrid models consistently outperformed simple single-stream architectures. The CNN+GRU, CNN 1D, and Fully-Connected DNN achieved the highest accuracy of 0.99, indicating their effectiveness in modeling the complex acoustic features associated with depression symptoms. Meanwhile, LSTM and GRU models, though competent in temporal learning, underperformed due to limited capacity in handling acoustic feature diversity, which may benefit more from convolutional encoding.

The combination of attention-based mechanisms with CNN and LSTM in the CNN+Transformer+LSTM architecture also showed high performance (accuracy 0.96), supporting previous findings that attention mechanisms can effectively capture global dependencies and highlight informative time segments in audio sequences. Overall, these results underline the significance of architectural design choices in achieving reliable depression classification from audio cues, particularly emphasizing the benefit of hybrid and attention-enhanced models.

Table 3. Evaluation per class label.

Models	Label	Precision	Recall	F1score	Accuracy
Fully-Connected DNN	Normal	0,97	1,00	0,99	0,99
	Depression Stage 1	1,00	1,00	1,00	
	Depression Stage 2	1,00	0,94	0,97	
CNN 1D	Normal	0,97	1,00	0,99	0,99
	Depression Stage 1	1,00	1,00	1,00	
	Depression Stage 2	1,00	0,94	0,97	
LSTM	Normal	0,69	0,97	0,80	0,76
	Depression Stage 1	0,77	0,38	0,51	
	Depression Stage 2	1,00	0,89	0,94	
GRU	Normal	0,61	0,97	0,75	0,71
	Depression Stage 1	0,83	0,19	0,31	
	Depression Stage 2	1,00	0,94	0,97	
CNN + Transformer	Normal	0,95	1,00	0,97	0,98
	Depression Stage 1	1,00	1,00	1,00	
	Depression Stage 2	1,00	0,89	0,94	
CNN + GRU	Normal	0,97	1,00	0,99	0,99
	Depression Stage 1	1,00	1,00	1,00	
	Depression Stage 2	1,00	0,94	0,97	
CNN + BiLSTM	Normal	0,95	1,00	0,97	0,98
	Depression Stage 1	1,00	1,00	1,00	
	Depression Stage 2	1,00	0,89	0,94	
CNN Attention Models + LSTM	Normal	1,00	0,97	0,99	0,96
	Depression Stage 1	0,90	1,00	0,95	
	Depression Stage 2	1,00	0,89	0,94	

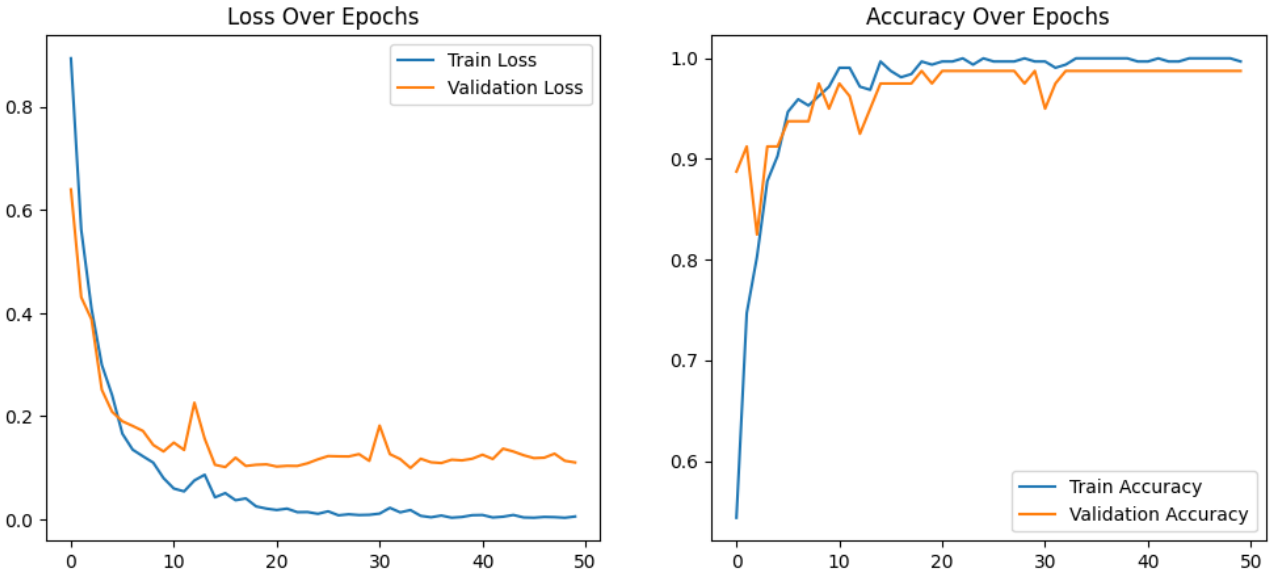

Fig. 2. Graphic loss and accuracy per epochs for model CNN-GRU.

4.2. Discussion

The experimental results demonstrate that hybrid deep learning architectures integrating convolutional and sequential components are effective in capturing discriminative acoustic patterns associated with depression severity in speech signals. Compared to standalone recurrent models such as LSTM and GRU, hybrid configurations—including CNN–GRU, CNN–BiLSTM, and CNN–Transformer–LSTM—consistently achieved higher performance across most evaluation metrics. This indicates that convolutional front-end layers play a critical role in extracting localized spectral and prosodic cues, which subsequently facilitate more informative contextual modeling. These findings are consistent

with prior studies [7, 12], which reported that CNN-based feature encoding significantly enhances downstream learning in speech-based affective computing tasks.

Despite these encouraging results, the near-perfect performance observed in several hybrid models must be interpreted with caution. The dataset used in this study comprises approximately 400 audio samples, which limits variability in speaker characteristics, recording conditions, and linguistic diversity. In low-resource settings, high accuracy values may reflect dataset homogeneity rather than genuine model generalization. Consequently, the reported results should be viewed as a proof-of-concept demonstration of architectural effectiveness rather than definitive evidence of robustness or clinical validity. Similar performance inflation has been reported in previous depression detection studies evaluated under constrained data regimes, where accuracy tends to decrease when tested on external or cross-corpus datasets.

Among the evaluated hybrid architectures, the CNN–Transformer–LSTM model achieved slightly lower overall accuracy than CNN–GRU but exhibited a more balanced F1-score across all depression severity classes. This suggests that self-attention mechanisms contribute to more stable class-wise discrimination by modeling global contextual relationships among acoustic descriptors. Attention-based representations allow the model to emphasize emotionally salient patterns that may be sparsely distributed within speech signals, which is particularly relevant for depression severity assessment [16, 17]. However, this representational advantage comes at the cost of increased computational complexity, highlighting an important trade-off between performance gains and deployment feasibility.

In contrast, simpler RNN-based models such as LSTM and GRU showed noticeably weaker performance, particularly in identifying Depression Stage 1. Mild depression often shares overlapping acoustic characteristics with both normal and severe states, making it inherently more challenging to distinguish. Without convolutional layers to capture fine-grained spectral variations, these models struggle to learn subtle affective differences from short-duration speech segments. Furthermore, the GRU model exhibited higher performance variance, indicating sensitivity to training data distribution—an issue that is exacerbated in small datasets. These observations align with prior work emphasizing the necessity of hierarchical and multi-scale feature representations for reliable speech-based mental health assessment [45].

When compared with existing state-of-the-art approaches summarized in Table 4, the present study is distinguished primarily by its focus on multiclass depression severity classification, whereas most previous works have concentrated on binary detection tasks [6, 10, 12]. Multiclass classification better reflects real-world clinical scenarios, where differentiating between mild and severe depression is critical for intervention planning. It is important to clarify that although EEG data are described as part of the dataset repository, the current experimental evaluation is restricted to audio-only modeling. Multimodal fusion is therefore positioned as a future research direction rather than a demonstrated contribution of this study.

Further insights were obtained through error analysis using confusion matrices. Misclassifications were most frequently observed between Depression Stage 1 and Stage 2, particularly in models lacking attention mechanisms. This suggests that fine-grained severity discrimination requires higher-level feature integration capable of capturing nuanced emotional and contextual patterns in speech. From a clinical perspective, confusion between mild and severe depression may result in delayed intervention or inappropriate treatment prioritization, underscoring the importance of severity-aware classification models.

Several limitations of the present study must be acknowledged. In addition to the small dataset size, speaker-independent partitioning could not be strictly enforced due to the absence of speaker identity annotations, which may introduce optimistic bias in performance estimates. A feature-level ablation study was not conducted, as the primary focus of this work was architectural benchmarking under low-resource conditions. Moreover, statistical significance testing and confidence interval estimation were not performed due to the single train–test split and limited data availability. Although model parameter counts are reported, inference latency, computational cost, and real-time deployment feasibility were not empirically evaluated and remain outside the scope of this study. As such, the findings should be interpreted as exploratory and preliminary rather than conclusive.

Future research will focus on expanding the dataset through larger, speaker-balanced, and multilingual speech corpora, as well as incorporating rigorous cross-validation and external benchmarking. The integration of explainable AI (XAI) techniques, such as SHAP or LIME, will be prioritized to enhance model transparency and interpretability. Additionally, exploring lightweight attention-based architectures may help balance predictive performance and computational efficiency, supporting more practical deployment scenarios.

Table 4. Comparative Summary with Related Studies.

Study	Modality	Deep Learning	Hybrid	Transformer	Attention	Classification Type	Acc.	Prec.	Recall	F1-Score
[7]	Audio	Yes	Yes	No	Yes	Binary	-	-	-	~089-0.91
[12]	Audio+Text	Yes	Yes	Yes	Yes	Binary	0.96	-	-	0.99
[3]	Audio+Text	Yes	Yes	No	Yes	Binary	0.76	-	0.86	0.82
[16]	Audio	Yes	Yes	No	Yes	Binary	0.81	0.65	0.80	0.70
Current Study	**Audio**	**Yes**	**Yes**	**Yes**	**Yes**	**Multiclass**	**0,99**	**0,98**	**0,99**	**0.99**

5. Conclusion

This study presented a systematic empirical evaluation of hybrid deep learning architectures for audio-based multiclass depression severity classification. By comparing combinations of CNN, GRU, BiLSTM, and Transformer components under a unified experimental setting, the study provided insights into how architectural design choices influence classification performance in low-resource speech scenarios. The results indicate that hybrid and attention-enhanced models can offer advantages over standalone architectures, particularly for severity-aware depression classification.

Nevertheless, the findings of this study should be interpreted with appropriate caution. The limited dataset size, absence of speaker-independent validation, and lack of external benchmarking constrain the generalizability of the reported results. Accordingly, the proposed framework should be regarded as an exploratory proof-of-concept rather than a clinically ready system.

Future work will focus on large-scale and longitudinal validation across diverse populations and languages, as well as the integration of multimodal information—including EEG, text, and visual cues—to improve robustness and interpretability. In addition, incorporating explainable AI techniques and ethical bias analysis will be essential to support responsible deployment in mental health screening contexts.

Author Contributions Statement

Neny Sulistianingsih – Conceptualization, Methodology, Supervision, Project administration, Writing – review & editing. Proposed the research direction, designed the overall experimental framework, supervised the implementation and evaluation process, and reviewed the manuscript for scientific rigor and clarity.

Galih Hendro Martono – Data curation, Software, Investigation, Model training, Validation, Formal analysis, Visualization, Writing – original draft. Conducted data acquisition and preprocessing, implemented the deep learning and hybrid architectures, executed model training and evaluation using standard metrics, prepared result visualizations, and drafted the initial manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

None.

Data Availability Statement

None.

Ethical Declarations

None.

Acknowledgments

None.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used generative AI–assisted tools exclusively to improve language quality, readability, and grammatical consistency, with full human oversight and critical review of all generated text. The use of AI was strictly limited to the writing process and did not affect the research design, data collection, analysis, model development, experimental evaluation, or interpretation of results. Generative AI tools were not used to generate scientific claims or conclusions, were not listed as authors or co-authors, and were not cited as responsible contributors. The authors remain fully accountable for the accuracy, integrity, and originality of the content presented in this work.

Abbreviations

The following abbreviations are used in this manuscript:

AI – Artificial Intelligence
CNN – Convolutional Neural Network
GRU – Gated Recurrent Unit
BiLSTM – Bidirectional Long Short-Term Memory
LSTM – Long Short-Term Memory
RNN – Recurrent Neural Network
Transformer – Transformer Neural Network Architecture
MFCC – Mel-Frequency Cepstral Coefficients
STFT – Short-Time Fourier Transform
MelSpec – Mel Spectrogram
DNN – Deep Neural Network
ML – Machine Learning
DL – Deep Learning
XAI – Explainable Artificial Intelligence
EEG – Electroencephalogram
MHA – Multi-Head Attention
BN – Batch Normalization
MP1D – 1D Max Pooling
TP – True Positive
TN – True Negative
FP – False Positive
FN – False Negative
PR-AUC – Area Under the Precision–Recall Curve
ROC-AUC – Area Under the Receiver Operating Characteristic Curve
NLP – Natural Language Processing

References

[1] G. H. Estimates, "Depression and Other Common Mental Disorders Global Health Estimates," 2017.
[2] S. S. Leal, S. Ntalampiras, and R. Sassi, "Speech-based Depression Assessment: A Comprehensive Survey," *IEEE Trans. Affect. Comput.*, vol. PP, no. 8, pp. 1–16, 2024, doi: 10.1109/TAFFC.2024.3521327.
[3] N. K. Iyortsuun, S. H. Kim, H. J. Yang, S. W. Kim, and M. Jhon, "Additive Cross-Modal Attention Network (ACMA) for Depression Detection Based on Audio and Textual Features," *IEEE Access*, vol. 12, no. January, pp. 20479–20489, 2024, doi: 10.1109/ACCESS.2024.3362233.
[4] D. K. Saha, T. Hossain, M. Safran, S. Alfarhood, M. F. Mridha, and D. Che, "Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-77193-0.
[5] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio-based depression detection using Convolutional Autoencoder," *Expert Syst. Appl.*, vol. 189, no. September 2021, p. 116076, 2022, doi: 10.1016/j.eswa.2021.116076.
[6] C. Sun, M. Jiang, L. Gao, Y. Xin, and Y. Dong, "A novel study for depression detecting using audio signals based on graph neural network," *Biomed. Signal Process. Control*, vol. 88, no. October 2023, 2024, doi: 10.1016/j.bspc.2023.105675.
[7] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomed. Signal Process. Control*, vol. 90, no. November 2023, p. 105898, 2024, doi: 10.1016/j.bspc.2023.105898.
[8] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-60278-1.
[9] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer++: A Hierarchical Efficient Framework for Paralinguistic Speech Processing," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 775–788, 2023, doi: 10.1109/TASLP.2023.3235194.
[10] Y. Sun, Y. Zhou, X. Xu, J. Qi, and F. Xu, "Weakly-Supervised Depression Detection in Speech Through Self-Learning Based Label Correction," *IEEE Trans. AUDIO, SPEECH Lang. Process.*, vol. 33, no., pp. 748–758, 2025, doi: 10.1109/TASLPRO.2025.3533370.
[11] J. Ye *et al.*, "Multimodal depression detection based on emotional audio and evaluation text," *J. Affect. Disord.*, vol. 295, no. February, pp. 904–913, 2021, doi: 10.1016/j.jad.2021.08.090.
[12] L. L. Gan, Y. Huang, X. Gao, J. Tan, F. Zhao, and T. Yang, "Multimodal Magic : Elevating Depression Detection with a Fusion of Text and Audio Intelligence."
[13] S. Yang *et al.*, "Fine-grained multimodal fusion for depression assisted recognition based on hierarchical knowledge-enhanced prompt learning," *Expert Syst. Appl.*, vol. 291, no. December 2024, p. 128532, 2025, doi: 10.1016/j.eswa.2025.128532.
[14] J. Ye *et al.*, "DEP-Former: Multimodal Depression Recognition Based on Facial Expressions and Audio Features via Emotional Changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2087–2100, 2024, doi: 10.1109/TCSVT.2024.3491098.

[15] L. Zhu *et al.*, "Explainable Depression Classification Based on EEG Feature Selection from Audio Stimuli," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 33, pp. 1411–1426, 2025, doi: 10.1109/TNSRE.2025.3557275.
[16] X. Xu, Y. Wang, X. Wei, F. Wang, and X. Zhang, "Attention-based acoustic feature fusion network for depression detection," *Neurocomputing*, vol. 601, 2024, doi: 10.1016/j.neucom.2024.128209.
[17] E. Lim, M. Jhon, J. W. Kim, S. H. Kim, S. Kim, and H. J. Yang, "A lightweight approach based on cross-modality for depression detection," *Comput. Biol. Med.*, vol. 186, no. September 2024, p. 109618, 2025, doi: 10.1016/j.compbiomed.2024.109618.
[18] L. Liu, S. Chen, K. Chen, J. Xu, and X. Chen, "Catching the Blackdog Easily: A Convenient Depression Diagnosis Method based on Audio-Visual Deep Learning," *IEEE Trans. Affect. Comput.*, vol. PP, pp. 1–16, 2025, doi: 10.1109/TAFFC.2025.3571697.
[19] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, and B. Alhaqbani, "High-Density Electroencephalography and Speech Signal Based Deep Framework for Clinical Depression Diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 20, no. 4, pp. 2587–2597, 2023, doi: 10.1109/TCBB.2023.3257175.
[20] F. Tian *et al.*, "Advancements in Affective Disorder Detection: Using Multimodal Physiological Signals and Neuromorphic Computing Based on SNNs," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 6, pp. 1–29, 2024, doi: 10.1109/tcss.2024.3420445.
[21] Y. Hu, J. Chen, J. Chen, W. Wang, S. Zhao, and X. Hu, "An Ensemble Classification Model for Depression Based on Wearable Device Sleep Data," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 5, pp. 2602–2612, 2024, doi: 10.1109/JBHI.2023.3258601.
[22] H. Fan *et al.*, "Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals," *Inf. Fusion*, vol. 104, no. July 2023, p. 102161, 2024, doi: 10.1016/j.inffus.2023.102161.
[23] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," *Comput. Speech Lang.*, vol. 55, pp. 40–56, 2019, doi: 10.1016/j.csl.2018.08.004.
[24] Hitler, "Multimodal dataset for depression analysis," *Kaggle*, 2025. https://www.kaggle.com/datasets/s3programmerlead/multimodal-dataset-for-depression-analysis (accessed Feb. 16, 2025).
[25] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017, doi: 10.1109/JSTSP.2017.2764438.
[26] K. Daly and O. Olukoya, "Depression detection in read and spontaneous speech: A Multimodal approach for lesser-resourced languages," *Biomed. Signal Process. Control*, vol. 108, no. May 2024, 2025, doi: 10.1016/j.bspc.2025.107959.
[27] E. Douglas-Cowie *et al.*, "Multimodal databases of everyday emotion: Facing up to complexity," *9th Eur. Conf. Speech Commun. Technol.*, pp. 813–816, 2005, doi: 10.21437/interspeech.2005-381.
[28] B. Schuller *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2001–2005, 2016, doi: 10.21437/Interspeech.2016-129.
[29] A. Khosla, P. Khandnor, and T. Chand, "Automated diagnosis of depression from EEG signals using traditional and deep learning approaches: A comparative analysis," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 108–142, 2022, doi: 10.1016/j.bbe.2021.12.005.
[30] B. McFee *et al.*, "librosa: Audio and Music Signal Analysis in Python," *Proc. 14th Python Sci. Conf.*, no. Scipy, pp. 18–24, 2015, doi: 10.25080/majora-7b98e3ed-003.
[31] M. Valstar *et al.*, "AVEC 2014 - 3D dimensional affect and depression recognition challenge," *AVEC 2014 - Proc. 4th Int. Work. Audio/Visual Emot. Challenge, Work. MM 2014*, no. January 2021, pp. 3–10, 2014, doi: 10.1145/2661806.2661807.
[32] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017, doi: 10.1109/LSP.2017.2657381.
[33] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," pp. 1–18, 2012, [Online]. Available: http://arxiv.org/abs/1207.0580.
[34] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, pp. 1–5, 2015, doi: 10.21437/interspeech.2015-1.
[35] S. Hochreiter and J. ¨urgen Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.
[36] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.
[37] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
[38] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," *2014 IEEE Glob. Conf. Signal Inf. Process. Glob. 2014*, pp. 577–581, 2014, doi: 10.1109/GlobalSIP.2014.7032183.
[39] A. Graves and J. Schmidhuber, "Framewise Phoneme Classi fi cation with Bidirectional LSTM Networks," *Neural Networks*, vol. 18, no. 5–6, 2005.
[40] N. Sulistianingsih and G. H. Martono, "Enhancing Predictive Models: An In-depth Analysis of Feature Selection Techniques Coupled with Boosting Algorithms," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 2, pp. 353–364, 2024, doi: 10.30812/matrik.v23i2.3788.
[41] S. Yoon and H. J. Yu, "BPCNN: Bi-Point Input for Convolutional Neural Networks in Speaker Spoofing Detection," *Sensors*, vol. 22, no. 12, pp. 1–21, 2022, doi: 10.3390/s22124483.
[42] J. Duque, F. Silva, and A. Godinho, "Data mining applied to knowledge management," *Procedia Comput. Sci.*, vol. 219, pp. 455–461, 2023, doi: 10.1016/j.procs.2023.01.312.
[43] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 1716–1720, 2018, doi: 10.21437/Interspeech.2018-2522.

[44] D. K. Sung, Y. Son, H. Eom, and S. Kim, "Improving I/O Performance via Address Remapping in NVMe Interface," *IEEE Access*, vol. 10, no. November, pp. 119722–119733, 2022, doi: 10.1109/ACCESS.2022.3221733.
[45] X. Lin, X. Zou, Z. Ji, T. Huang, S. Wu, and Y. Mi, "A brain-inspired computational model for spatio-temporal information processing," *Neural Networks*, vol. 143, no., pp. 74–87, 2021, doi: 10.1016/j.neunet.2021.05.015.

Authors' Profiles

Neny Sulistianingsih received a Bachelor of Informatics (S.Kom) and a Master of Computer (M.Kom) from the Universitas Islam Indonesia, Yogyakarta. She completed her Doctoral studies in 2023 at the Universitas Gadjah Mada, Yogyakarta. Her research interests include Machine Learning, Data Mining, Graph Mining, Big Data, Social Network Analysis, and Natural Language Processing. She is currently a Lecturer at Bumigora University.

Galih Hendro Martono received a Bachelor of Informatics (S.Kom) from Universitas Islam Indonesia and a Master of Engineering from Universitas Gadjah Mada, Yogyakarta. He completed her Doctoral studies in 2023 at the Universitas Gadjah Mada, Yogyakarta. Her research interests include Machine Learning, Data Mining, Graph Mining, Big Data, Social Network Analysis, and Natural Language Processing. He is currently a Lecturer at Bumigora University.

How to cite this paper: Neny Sulistianingsih, Galih Hendro Martono, "A Robust Hybrid Deep Learning Model for Multiclass Depression Classification from Speech Audio", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.2, pp. 124-136, 2026. DOI:10.5815/ijigsp.2026.02.08