Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 77-85. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2742

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AN ENHANCED MULTIMODAL APPROACH FOR SENTIMENT ANALYSIS USING DEEP LEARNING TECHNIQUES

VinithaV[1]* and S.K. ManjuBargavi[2]

[1]Research scholar, DepartmentofComputerScienceandIT,Jain(Deemed-to-be-University),Bengaluru,India
[2]Professor, Department of Computer Science and IT, Jain(Deemed-to-be-University),Bengaluru,India

[1]https://orcid.org/0009-0006-4586-5923, [2]https://orcid.org/0000-0001-7112-5810

Email: * vinithafive@gmail.com, b.manju@jainuniversity.ac.in

ARTICLE INFO

Article History
Received: September 25, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Multimodal,
Hybrid remoraoptimization,
Depression detection,
Sentiment analysis,
Deep learning.

ABSTRACT

Social media has become a prominent medium through which individuals express their perspectives, sentiments, and experiences, frequently exhibiting apparent indications of mental health conditions such as depression. The identification of depression through nonverbal behaviour has garnered considerable interest. Implementing previous research concerning the detection of depression within real-world scenarios presents challenges, primarily because the research predominantly concentrated on identifying depressive individuals within controlled laboratory settings. Depression affects millions worldwide therefore, early and accurate detection is critical for prompt intervention. Conventional diagnostic methodologies are often subjective in nature and require considerable time to implement. This study introduces a hybrid remora-optimized multimodal deep learning (HRO-MDL) approach to incorporate multi-modal data such as audio, video, and text for enhanced performance in the assessment of depressive emotional states and sentiments. The proposed framework accurately predicts sentiment across modalities by integrating feature extraction from various modalities using a deep convolutional neural network optimized with hybrid optimization. The model's performance was assessed using the D-Vlog and Depression Cleaned Reddit datasets, demonstrating notable accuracy levels 94.00% for text data, 92.00% for audio data, and 93.00% for video data. In contrast to traditional methodologies, the presented work exhibits enhanced efficiency and underscores its suitability for diverse multimodal sentiment analysis applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The significant influence of mental health on both individual and societal well-being has resulted in increased focus on this issue among diverse populations. Mental health disorders, such as depression, are commonly occurring conditions that drastically impact millions of people worldwide. The disease results with constant pain, deep despair and considerably impaired functional capacity that leads to serious limitation in daily life [1]. Current method for diagnosing depression is primarily self-reported questionnaires, interviews, and clinical measurements. However, these approaches can be costly, time-consuming and subject to human perception bias [2]. To assess people's emotional states, clinicians are using modality data including text, audio and visual [3]. An accurate and real-time detection can be realized if the analysis is considering multiple depression data sources including text, voice, and visual to learn depression more explicitly [4]. The combination of multiple modalities in such systems is helpful in improving result of confidence, as it uses the strengths of different signals that reduces the chances of signal failure and misinterpretation. The approach maintains the continuous monitoring with increased precision with effect of early detection [5] and even in the analysis of large data such as text, audio and video data, there are various more advanced models (such as CNN's, RNN's and transformers) showing increased effectiveness. Medical professionals can leverage these models to identify signs of depression, which may present as subtle behaviours that includes inclination toward the condition. The methods for detecting depression support the advancement of deep learning technologies, enabling precise and

early identification of depressive symptoms [6]. Current state-of-the-art approaches, including CNNs, RNNs, and transformers, demonstrate robust performance in analysing multimodal data comprising text, audio, and video that involves substantial textual content. The model assists healthcare providers in identifying the symptoms of depression and detects subtle behavioural changes indicative of a tendency toward depressive patterns [7]. The primary challenge is to improve model efficiency while maintaining generalizability across diverse populations with varying linguistic and socioeconomic attributes [8].

This method offers a significant degree of accuracy independently however, the necessity for these models to be resilient and effective leads to overfitting due to insufficient real-world applicability [9]. This study explores a multi-modal deep learning approach incorporating hybrid remora- Integrated Grey Wolf Optimizationtechniques. This method is structured to preprocess various types of input data (text, audio, video) through multiple stages, including data preprocessing and subsequently proceeding to multimodal feature extraction. DenseNet-201 classification is utilized to predict emotional states of depression and sentiments, as it is well-suited for diverse input modalities. Additionally, this optimization technique is employed to refine the models and improve system performance. The hybrid optimization enhances system accuracy, rendering it robust for practical applications. Below is the major contribution of the proposed approach:

- To develop a multi-modal deep learning approach with hybrid remora optimization and to predict the emotional state of depression and sentiments across various modalities.
- The proposed model processes the sentiment of each modality (text, acoustic, visual) by modelling the verbal and non-verbal behaviour over the modalities.
- The Hybrid Remora-Integrated Grey Wolf Optimizationwith the DenseNet-201 approach is utilized to compare with current baseline models and achieve the best feature set with the highest accuracy.

II. LITERATURE REVIEW

Depression is one of the most prevalent and severe mental health disorders identified on social media. The development of an attention-based approach for detecting depression. This model emphasized the most important features of the input data and increased its interpretability. The observed overfitting was due to the model's increasing complexity [10]. Spatial attention networks are created to detect depression in facial images. The approach demonstrated significant efficacy and improved the detection parameters. The dataset for this model was significantly smaller and more complex [11]. Four hybrid deep learning models were created specifically to identify symptoms of melancholy. This method may function as an efficient instrument for detecting symptoms of depression on social media platforms, owing to its improved accuracy and relevance to real-world textual data. This method largely analysed textual data, overlooking potentially beneficial multimedia data such as images or video snippets, that could offer a more comprehensive understanding of depression [12], [13]. They devised an attention technique that amalgamated voice and text analysis to detect depression. This method derives contextual dependencies from both prior and later segments of the sequence by analysing data in both forward and reverse directions.

The bidirectional data processing incurred computational costs and led to an overfitting problem [14]. A fusion fuzzy model was used to identify depression from facial expressions. For visual data, this approach demonstrated strong feature extraction capabilities, but it becomes challenging when working with large datasets [15]. The deep learning model performed better at detecting depression in terms of accuracy rate and using data from Twitter. The model is susceptible to overfitting to the specific dataset, which could restrict its effectiveness when applied to newly obtained, unseen data [16]. A method for identifying depression and capturing local dependencies as well as sequential relationships within the text was introduced using deep learning. By capturing the semantic substance of sentences not explicitly present in the lexicon, this approach enhanced the model's ability to interpret subtle verbal indicators of depression. It required a substantial amount of computational power, especially when handling big datasets [17].

Experimental outcomes emphasize the advantages of integrating deep learning with multi-aspect data, indicating that it exceeds several established and robust baseline methods [18]. The Particle Swarm Optimization algorithm was developed for the diagnosis of depression. This approach enhances both precision and the rate of convergence. The accuracy of the model is significantly influenced by the quality and diversity of the dataset. Imbalanced datasets can negatively impact the model's performance, resulting in inaccurate melancholy detection, despite the attention given to recognizing melancholy in social media [19-21]. Their research has enhanced the model's interpretability and demonstrated the significance of features within the input data. Due to the increased complexity of the models, overfitting occurred. For the purpose of detecting depression in facial images, this study employed spatial attention networks [22], [23]. The primary benefits of this methodology were its enhanced performance characteristics and the improved pertinence of the generated features for detection purposes. The resultant model was both decreased in quantity and increased in complexity [24].

III. PROPOSED METHODOLOGY

This study focuses on the Hybrid RemoraOptimizedMultimodal Deep Learning (HRO-MDL) approach which is proposed to predict the emotional state of depression and sentiment detection. Depression detection through the integration of multi-modal data (video, audio, and text) utilizing different methods. It encompasses a variety of phases, including data preprocessing, parameter tuning, classification, and optimization, as illustrated in Figure 1.

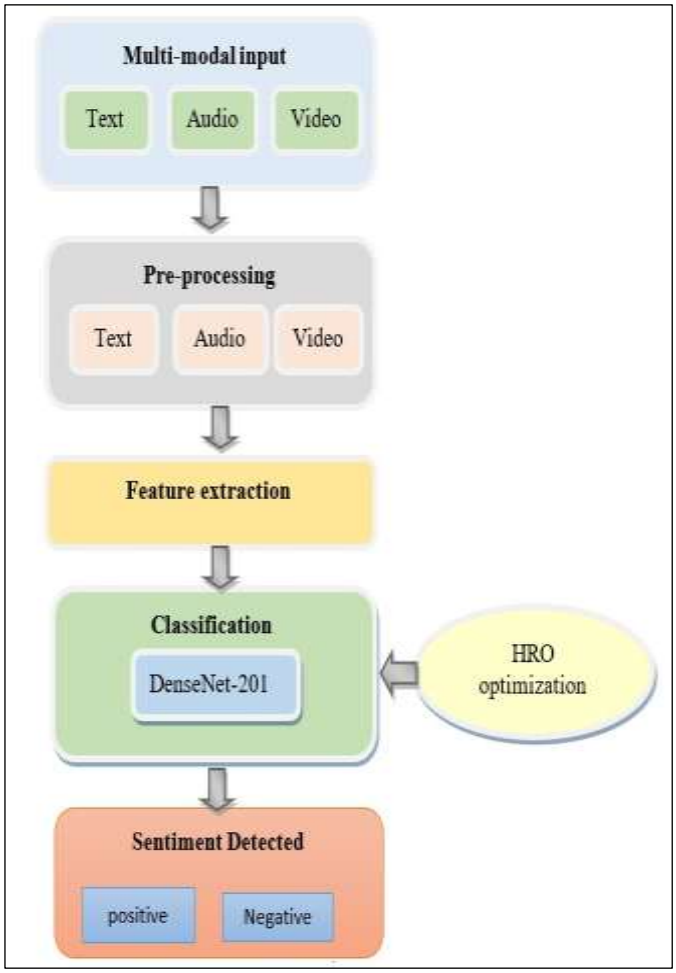

Figure 1: Proposed Multi-Modal Framework.
Source: Authors, (2026).

III.1 DATA COLLECTION

The input source audio and video data are collected from the D-Vlog dataset. The textual dataset is taken from Cleaned Reddit dataset. The audio features that supply important cues to diagnose depression, such as facial expressions, speech patterns, and emotional response signals, are also taken from the D-Vlog dataset, which also includes video logs of human life experiences. It consists of depression-related linguistic patterns as well as sentiment indicators and psychological markers for depression assessment. The mathematical representation of multi-modal input is represented as follows.

The D-Vlog dataset is the source of the input audio and video data. Cleaned Reddit depression dataset used for textual data. Also extracted D-Vlog datasetincludes video records of human life experiences, audio elements that provide crucial clues to diagnose depression, such as speech patterns, facial expressions, and emotional response signals. It includes sentiment indicators, psychological markers for depression assessment, and linguistic patterns associated with depression. Multi-modal input can be represented mathematically as follows:

$$Y = \{Y_T, Y_A, Y_V\} \tag{1}$$

Whereas, audio and video data from the D-Vlog dataset is represented byY_A,Y_V, and text data denoted as Y_T from the Reddit Cleaned depression dataset.

III.2 PRE PROCESSING

Preprocessing of audio, video, and textual data is utilized to extract multimodal features for sentiment prediction and depression detection. Speech is pre emphasized, framed, and converted to the Mel Frequency Cepstral Coefficients (MFCC), which are used to measure speechand ambient clues. Videos are converted into frames, resized, normalized, and spatially featured positions captured by the CNNs and temporally motion patterns learned by optical flow. Text data preprocessing, involving tokenization, lemmatization, stop word removal, and embedding generation, is conducted to preserve the semantics and context of the text data for analysis. Let the text input be expressed in Equation 2.

$$Y_T = \{m_1, m_2, \ldots . m_n\} \tag{2}$$

Where, mi represents the individual sentence

III.3 FEATURE EXTRACTION

The Audio, visual, and text are all used to extract multimodal features to model sentiment detection using the Dvlog dataset and Reddit dataset. Then texts are pre-processed through tokenization and cleaning and are transformed into contextual embeddings based on XLNet, which can obtain deep semantics and syntax relations. Also, the audio streams are pre emphasized, framed, and converted to Mel Frequency Cepstral Coefficients (MFCC) suitable for time sequence modelling with XLNet on temporal acoustic patterns. Videos are split into frames and resized, and then normalize, spatial features are extracted with CNNs (e.g., DenseNet), and temporal dependencies models such as XLNet transformers with optical flow for motion dynamics. Given a sequence of input tokens $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_T}]$, XLNet models the joint probability of a given input sequence of tokens as per Equation 3.

$$P(y) = \prod_{t=1}^{T} P(y_t | y_{<t}) \tag{3}$$

The set of tokens that have been proceeding in the permutation is depicted. XLNet's ability to analyse all sequence permutations, which makes each token accessible to all other tokens, can be interpreted to serve as an alternative to the traditional prediction approach systems, where tokens are only conditioned on a fixed left-to-right context through an autoregressive model. As a result, it achieves bidirectional dependency understanding. The anticipated log likelihood is optimized by the model for each possible sequence of manufacturing order permutations as per Equation 4

$$\hat{\theta} = arg\, \max_{\theta} E_{Z\sim\Im}\left[\sum_{t=1}^{T} log\, P\left(y_{Z(t)} | y_{z(<t)}\right)\right] \tag{4}$$

III.4. CLASSIFICATION

Once that features are gathered, they are used as input to the DenseNet-201 deep learning model for classification, which includes efficient dense connections elements that includes feature reuse but also improves gradient flow. The applicability of the model parameters is enhanced through Hybrid Remora-Integrated Grey Wolf Optimization with the faster converging speed, and the highest classification accuracy. The collaboration of extraction methods and optimized classification techniques leads to solid outcomes in both depression evaluations and sentiment analysis processes. Each layer of Dense Net- 201 is based on the dense block concept, Ds receive the output of all preceding layers 〖D〗_1,D_2,....D_(s-1). The output of the s-th layer in a dense block can be represented as follows for a particular input feature map as per Equation 5.

$$Y_s = H_s([Y_0, Y_1, \ldots Y_{s-1}]) \tag{5}$$

Where, the output of the s-th layer is indicated by Y_s, while the function is shown by $H_s$applied at layer s,$[Y_0, Y_1, \ldots Y_{s-1}]$ representing the concatenation of all previous layer output. DenseNet-201 runs several dense blocks, transition layers, and a fully linked layer for classification to provide its final output. The SoftMax function is used to generate the output probability vector at the end of the classification process represented in Equation 6.

$$P(x|Y) = soft\, max(WY_T + b) \tag{6}$$

Where Y_T denotes the final feature map after being passed through all layers, W and b are weights and bias of the fully connected layer,$P(x|Y)$ denotes the probability distribution over classes after the softmax activation. The usage of a SoftMax function for classification and DenseNet-201's dense connection are reflected in the above formulation.

III.5 HYBRID REMORA OPTIMIZATION ALGORITHM

It is a kind of population based metaheuristic optimization algorithm that imitates hierarchical hunting strategy of Grey Wolf Optimization (GWO) for group hunting and embodies the symbiotic follow-and-adapt behaviour of remora fish in oceanic ecosystems. Within this hybrid framework, the leadership hierarchy of Grey Wolf serves as the primary search approach, steering global exploration of solution space. The follow-and-adapt mechanism as a biologically inspired approach underlies remora integration strategyand serves as an effective mechanism for both exploration and exploitation balance and thus enhancing convergence speed and solution quality. The positions of remora agents will be updated in each iteration according to the associated position update formulation defined in Equation (7) during the remora adaptation phase.

$$Y_i(t+1) = Y_{best} + \beta(Y_{best} - Y_i(t)) \tag{7}$$

The new position of remora i is represented as Yi(t+1), the optimal solution is indicated as Y_best , and the adaptive parameter governing fine-tuning is denoted as β . The Eat Thoughtfully Phase enables remoras to leverage possible solutions, resulting in accelerated convergence and enhanced accuracy, particularly in complex optimization challenges such as hyperparameter tuning in deep learning. However, it achieves better exploration and exploitation balance through an integration with RSO. The purpose behind this modification serves to stop early convergence while simultaneously enabling better global exploration and effective optimization of deep learning models. This proposed method enables superior performance with optimal accuracy and robustness.

IV. RESULTS AND DISCUSSION

This section focuses on assessing the performance of the proposed system and comparing it against baseline approaches using standard benchmark datasets. To make the model more robust, 2 standard datasets (DVLOG, CLEANED REDDIT), are used to integrate to form a hybrid model.

IV.1. DATASET DESCRIPTION

- D-Vlog Dataset [19]: It consists of 961 YouTube vlogs (160 hours) used for depression detection using non-verbal behaviour. The dataset contains 816 subjects (406 controls, 555 sufferers of depression) with audio and visual features saved as NumPy files. This supports the automated depression detection in research with AI-based analysis of facial expressions, gestures, and speech patterns as they naturally occur in everyday life.
- DepressionReddit Dataset: This dataset contains 7,000 labelled text data that are collected from depression related subreddits. Text data reflects real emotional and mental health conditions through self-expressed content. It is valuable for sentiment and depression detection, offering linguistic markers that help identify psychological distress using natural language processing techniques.

IV.2. EXPERIMENTAL SETUP

In the context of this study, where training, testing is done, after was 70% data trained and other thirty percent is imposed to test. The study done on a 2.4 GHz Intel(R) processor with 16 gigabytes of random-access memory (RAM).

IV.3. EVALUATION METRICS

Several evaluation metrics were used to assess the global classifying efficiency. Table 1 was used to evaluate the proposed approach for different evaluation criteria to analyse the sentiments with various performance metrics as shown in Table 1 and Fig.4.

IV.4. CONFUSION MATRIX HEATMAP

It denotes the evaluation of a classification model. It shows real classifications versus predictions in a matrix arrangement across modalities shown in below Figure 2, the confusion matrix shows how well system predict accurate and inaccurate predictions across depression vs normal classes over different attributes. Darker cells correspond to high-accuracy predictions, and lighter cells are misclassifications. In Figure 3. Shows Training and validation accuracy and loss curve for various modalities, Figure 4 shows ROC Curve across thresholds on classifier performance, AUC of 0.95–0.98 indicates that our sentiment classification is highly sensitive and specific.

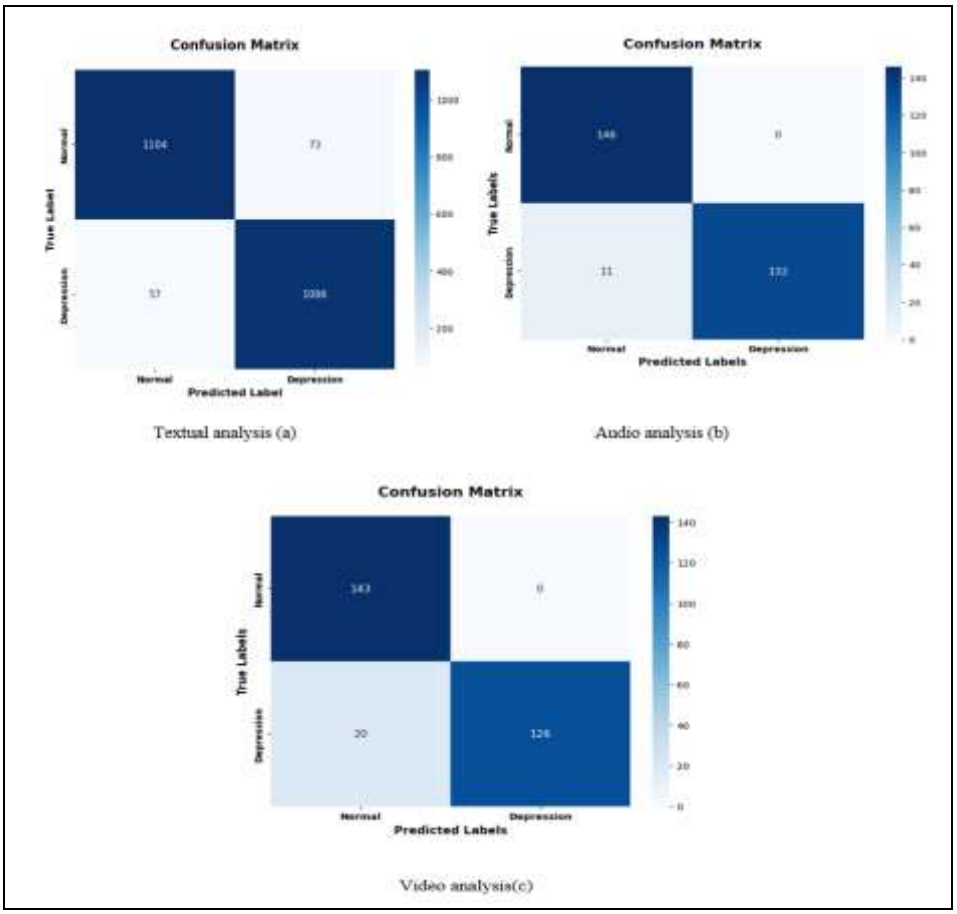

Figure 2: Confusion matrix heat map for text, audio, video modalities.
Source: Authors, (2026).

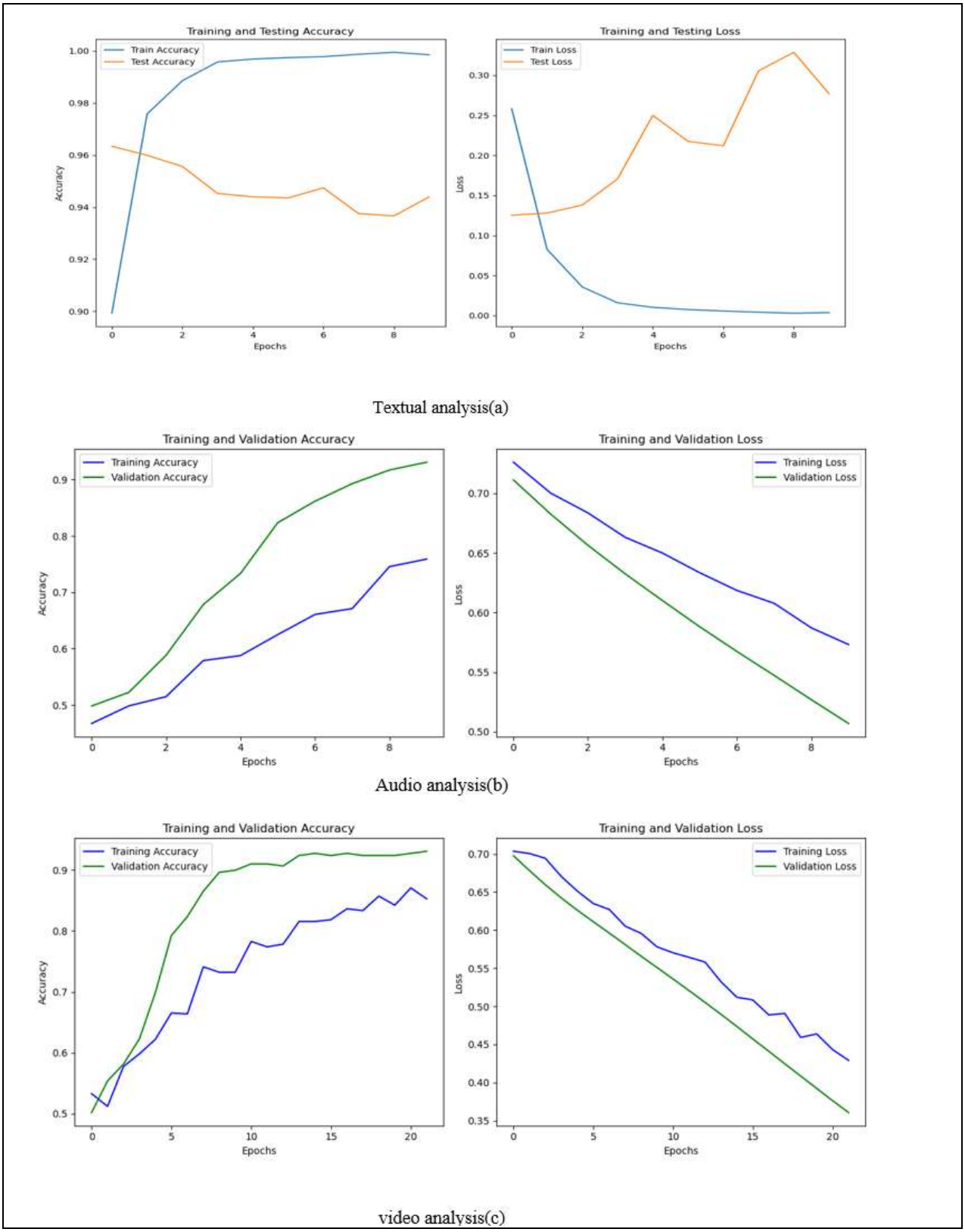

Figure 3: Training and Validation Accuracy and loss acrossmodalities.
Source: Authors, (2026).

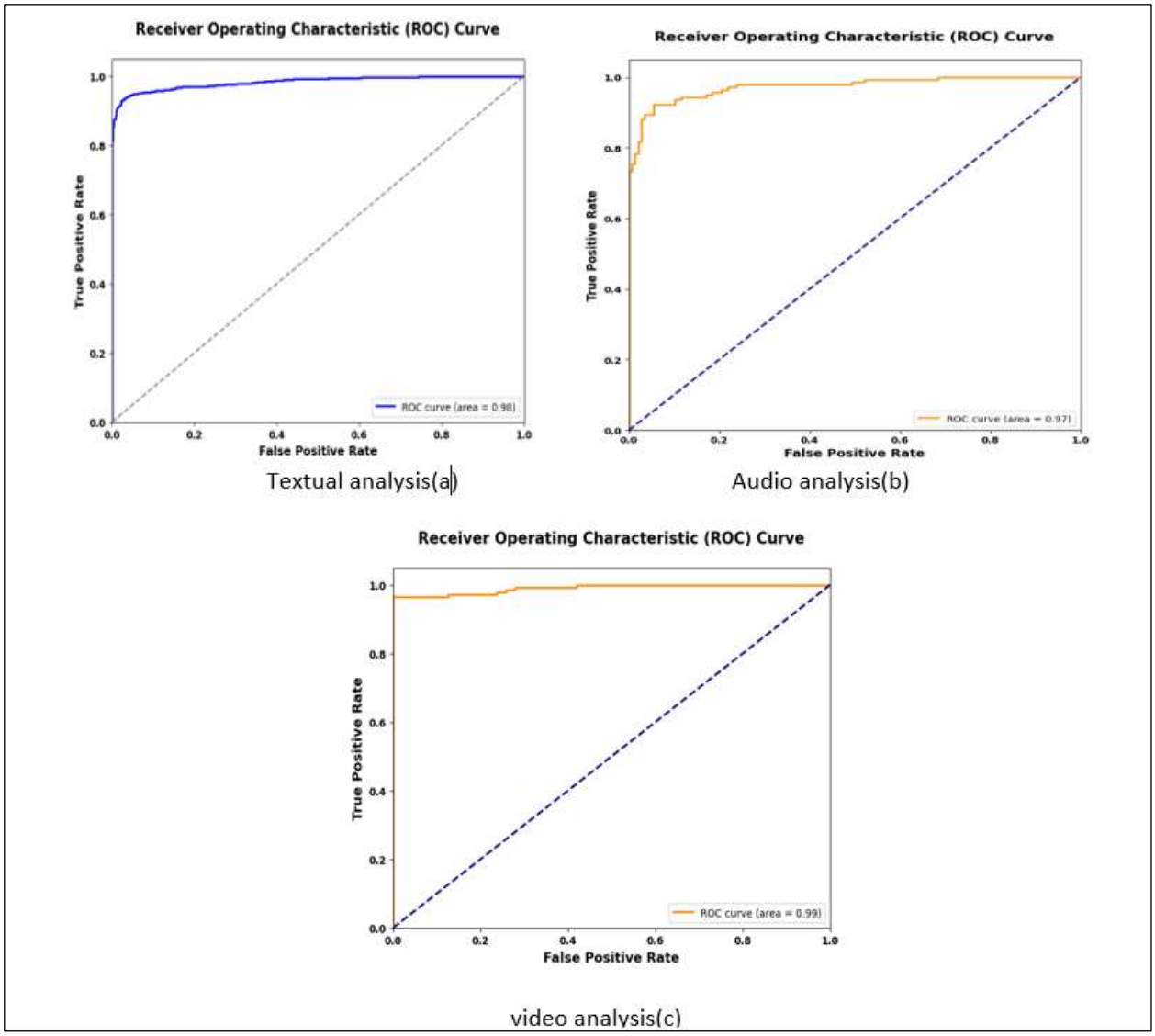

Figure 4: ROC for various modalities.
Source: Authors, (2026).

IV.5. COMPARATIVE STUDY WITH VARIOUS BASELINE MODELS

The proposed depression detection system is a webbasedapplication that contains multiple interactive modulestoexecute regular mental health assessments. The Base line models such as Tensor fusion Network (TFN), Depression Detector, Time-Aware Attention Multimodal Fusion Network (TAMFN), STST Spatiotemporal Semantic Transformer (STST) were tested over problems in depression detection like inability of handling sequential data, high computation complexity and slower execution on very large datasets shown by Table 1 and Figure 5. However, these methods consistently underperform due to issues related to hyperparameters, overfitting, and convergence. Issues can be addressed through integrationof Hybrid Remora-Integrated Grey Wolf Optimization, usingDenseNet-201 for classification and Adam optimizer was used for fine tuning to improve the performance. It also provides a real time facial landmark detection, contributes to depression assessment and there other several use cases. The different baseline models are discussed below.

1. **TFN (Tensor Fusion Network):** This is a multimodal sentiment baseline that models unimodal, bimodal and trimodal interactions explicitly through tensor outer-products, it's strong but it has many parameters, it works poorly when the data is very sparse.
2. **Depression Detector:** Audio + Text bi-modal depression detector with GRU/BiLSTMmodel for temporal modeling, which is strong base-midline for clinical/interview-style data.
3. **TAMFN:** Time-Aware Attention Multimodal Fusion Network is used for learning the temporal salience and cross-modal fusion for depression detection which deals with the temporal handling issue better than TFN.
4. **STST:** Spatiotemporal Semantic Transformer uses attention over visual/text/audio streams with semantic alignment, improved long-range temporal modeling and fusion.
5. **TVLT**:(textless pretraining) token-based vision-language transformer, re-purposed for cross-modal understanding and strong generic prior, moderate domain gap without task-specific pretraining.

Table 1: Comparison across various baseline models.

Model Type	Precision	Recall	F1-Score	Acc
TFN [18]	61.39%	62.26%	61.00%	61.55%
Depression Detector [19]	65.40%	65.57%	63.50%	64.82%
TAMFN [20]	66.02%	66.50%	65.82%	66.11%
STST [21]	72.50%	77.67%	75.00%	75.06%
TVLT [22]	67.30%	68.03%	67.80%	67.80%
Proposed Model				
Text Analysis	94.00%	95.00%	94.00%	94.00%
Audio Analysis	93.00%	91.00%	92.00%	92.00%
Video Analysis	94.00%	92.00%	93.00%	93.00%

Source: Authors, (2026).

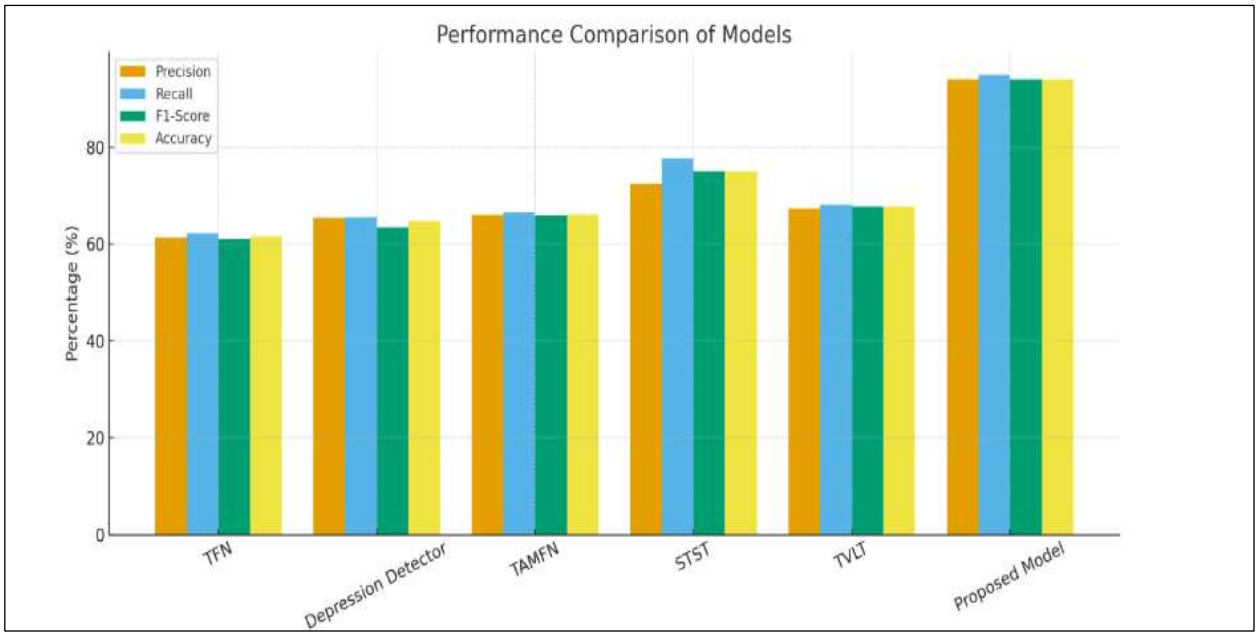

Figure 5: Comparison graph with various Baseline Models.
Source: Authors, (2026).

V. CONCLUSION

Despite the field of deep learning is constantly expanding, there is still a lot of work accessible for depression identification. This study proposes to use multimodal data (text, video, and audio) that is effectively incorporated to detect emotional states of depression and classify sentiment. High-level features between modalities can then be extracted using feature extraction techniques. The Dense Net-201 model is then used to achieve an accurate classification, and a Hybrid Remora-Integrated Grey Wolf OptimizationAlgorithm has been introduced to further improve the classification accuracy. Initially, it preprocesses the various input data so that it can support text, audio, and video data separately with the available data. Thus, it has been demonstrated that this suggested method outperforms current methods in the comparative analysis, with greater accuracy rates for text data with 94%, audio data is 92%, and video data is 93%. The integration of real-time sensor physiology will be the focus of future research in order to create a more interpretable visualization of the information that the mental health professionals are actually acquiring from the decisions being performed.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Vinitha V, S.K. Manju bargavi.
Methodology: Vinitha V, S.K. Manju bargavi.
Investigation: Vinitha V, S.K. Manju bargavi.
Discussion of results: Vinitha V, S.K. Manju bargavi.
Writing –Original Draft: Vinitha V.
Writing –Review and Editing: Vinitha V, S.K. Manju bargavi.
Resources: Vinitha V, S.K. Manju bargavi.
Supervision: Vinitha V, S.K. Manju bargavi.
Approval of the final text: Vinitha V, S.K. Manju bargavi.

VII. ACKNOWLEDGMENTS

The authors would like to extend their heartfelt thanks to the Department of Computer science and IT, Jain University, Bengaluru, for their support of this work.

VIII. REFERENCES

[1] S. Khan and S. Alqahtani, “Hybrid machine learning models to detect signs of depression,” Multimedia Tools and Applications, vol. 83, no. 13, pp. 38819–38837, 2024.

[2] K. Jawad et al., “Novel cuckoo search-based metaheuristic approach for deep learning prediction of depression,” Applied Sciences, vol. 13, no. 9, p. 5322, 2023.

[3] N. Marriwala and D. Chaudhary, “A hybrid model for depression detection using deep learning,” Measurement: Sensors, vol. 25, p. 100587, 2023.

[4] A. S. Rajawat et al., “Fusion fuzzy logic and deep learning for depression detection using facial expressions,” Procedia Computer Science, vol. 218, pp. 2795–2805, 2023.

[5] H. A. Sanghvi et al., “A deep learning approach for classification of COVID and pneumonia using DenseNet-201,” International Journal of Imaging Systems and Technology, vol. 33, no. 1, pp. 18–38, 2023.

[6] Y. Pan et al., “Spatial–temporal attention network for depression recognition from facial videos,” Expert Systems with Applications, vol. 237, p. 121410, 2024.

[7] W. Zhang, J. Xie, Z. Zhang, and X. Liu, “Depression detection using digital traces on social media: A knowledge-aware deep learning approach,” Journal of Management Information Systems, vol. 41, no. 2, pp. 546–580, 2024.

[8] D. K. Saha et al., “Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks,” Scientific Reports, vol. 14, p. 25470, 2024.

[9] N. K. Iyortsuun et al., “Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features,” IEEE Access, vol. 12, pp. 20479–20489, 2024.

[10] A. S. Rajawat et al., “Fusion fuzzy logic and deep learning for depression detection using facial expressions,” Procedia Computer Science, vol. 218, pp. 2795–2805, 2023.

[11] D. S. Khafaga et al., “Deep learning for depression detection using Twitter data,” Intelligent Automation & Soft Computing, vol. 36, no. 2, pp. 1301–1313, 2023.

[12] V. Tejaswini, K. S. Babu, and B. Sahoo, “Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, pp. 1–20, 2024.

[13] A. F. Dharma, “A deep learning using DenseNet201 to detect masked or non-masked face,” 2021.

[14] H. Zogan et al., “Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media,” World Wide Web, vol. 25, no. 1, pp. 281–304, 2022.

[15] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, “An attention-based CNN-BiLSTM model for depression detection on social media text,” Expert Systems with Applications, vol. 249, p. 123834, 2024.

[16] Y. Pan et al., “Spatial–temporal attention network for depression recognition from facial videos,” Expert Systems with Applications, vol. 237, p. 121410, 2024.

[17] M. Kächele, M. Schels, C. Thiel, and F. Schwenker, “Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression,” in Proc. 6th Int. Workshop on Audio/Visual Emotion Challenge, pp. 11–18, 2016.
[18] J. Yoon, C. Kang, S. Kim, and J. Han, “D-vlog: Multimodal vlog dataset for depression detection,” in Proc. AAAI Conf. Artif. Intell., 2022.

[19] L. Zhou et al., “TAMFN: Time-aware attention multimodal fusion network for depression detection,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 669–679, 2022.

[20] Y. Tao et al., “Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer,” Digital Communications and Networks, vol. 10, no. 3, pp. 577–585, 2024.

[21] P. Moon and P. Bhattacharyya, “We care: Multimodal depression detection and knowledge infused mental health therapeutic response generation,” arXiv preprint arXiv:2406.10561, 2024.

[22] K. S. Seby, M. Elamparithi, and V. Anuratha, “A hybrid attention-based deep learning system for suicidal ideation detection in social media,” 2024.

[23] E. Yeskuatov, S. L. Chua, and L. K. Foo, “Detecting suicidal ideations in online forums with textual and psycholinguistic features,” Applied Sciences, vol. 14, no. 21, p. 9911, 2024.

[24] S. L. Mirtaheri, S. Greco, and R. Shahbazian, “A self-attention TCN-based model for suicidal ideation detection from social media posts,” Expert Systems with Applications, vol. 255, p. 124855, 2024.