SOCIALMIND: MULTIMODAL FUSION AND TEMPORAL MODELING FOR MENTAL HEALTH PREDICTION FROM SOCIAL MEDIA DATA

Dr. R. RAJA SEKHAR
Professor
Department of Computer Science and Engineering
JNTUA College of Engineering (Autonomous)
Ananthapuramu, Andhra Pradesh, India
drasharaj2002.cse@jntua.ac.in

YEDDULA TEJASWINI
PG Scholar
Department of Computer Science and Engineering
JNTUA College of Engineering (Autonomous)
Ananthapuramu, Andhra Pradesh, India
tejaswiniyeddula2@gmail.com

ABSTRACT

Mental health disorders such as Depression, Anxiety, Bipolar Disorder, and Attention Deficit Hyperactivity Disorder (ADHD) affect millions of people worldwide and require timely identification for effective intervention. Traditional diagnostic approaches often depend on clinical evaluations, which may not provide continuous monitoring of an individual's psychological condition. With the increasing use of social media, large volumes of user-generated text, images, and audio content offer valuable insights into emotional and behavioral states. This research proposes TriModalSenseNet, a novel multimodal deep learning framework for early mental health disorder detection using social media data. The framework integrates textual, visual, and audio modalities to capture comprehensive behavioral characteristics. Text features are extracted using BERT, image features through ResNet-50, and audio representations using wav2vec 2.0. An Attention-Based Deep Neural Network performs multimodal feature fusion, while a Long Short-Term Memory (LSTM) network models temporal behavioral changes over time. Experimental results demonstrate the effectiveness of the proposed approach in classifying ADHD, Anxiety, Bipolar Disorder, and Depression. The model achieved 92.60% accuracy, 91.60% precision, 91.20% recall, and 91.00% F1-score, with the best performance obtained at Epoch 20. The findings indicate that multimodal learning combined with temporal analysis significantly enhances mental health prediction performance, supporting intelligent healthcare systems for early detection and intervention.

Keywords: Mental Health Detection, TriModalSenseNet, Multimodal Learning, BERT, ResNet-50, wav2vec 2.0, Attention Network, LSTM, Social Media Analytics.

1. INTRODUCTION

Mental health disorders have emerged as one of the most significant public health challenges of the twenty-first century, affecting individuals across all age groups, socioeconomic backgrounds, and geographical regions. Conditions such as depression, anxiety, bipolar disorder, and Attention Deficit Hyperactivity Disorder (ADHD) can substantially impact an individual's emotional well-being, cognitive functioning, and quality of life. According to global health reports, mental health disorders contribute significantly to the overall burden of disease and disability worldwide, emphasizing the need for timely diagnosis and intervention [1].

The rapid growth of social media platforms has transformed the way individuals communicate, express emotions, and share personal experiences. Platforms such as Twitter, Reddit, Facebook, and Instagram generate vast amounts of user-generated content that reflects psychological states, behavioral patterns, and emotional responses. Consequently, social media data has become a valuable resource for researchers seeking to understand and predict mental health conditions through computational methods [2].

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have enabled the automated analysis of large-scale social media data for mental health assessment. Existing studies have utilized traditional machine learning algorithms, ensemble learning approaches, and Large Language Models (LLMs) to identify indicators of mental health disorders from textual content. Although these approaches have demonstrated promising results, they primarily focus on textual information and often fail to capture the rich emotional and behavioral signals

available in visual and audio modalities. Furthermore, many existing systems lack the capability to analyze temporal changes in user behavior, which are crucial for detecting the gradual onset and progression of mental health conditions.

Human emotions and psychological states are inherently multimodal, being expressed through language, facial expressions, voice characteristics, and behavioral interactions. Therefore, integrating multiple data modalities can provide a more comprehensive understanding of an individual's mental state. Multimodal learning techniques have shown significant potential in extracting complementary information from heterogeneous data sources, leading to improved predictive performance and robustness in affective computing applications.

To address the limitations of existing approaches, this research proposes a novel multimodal deep learning framework for early mental health disorder detection. The proposed system integrates textual, visual, and audio information using Bidirectional Encoder Representations from Transformers (BERT), ResNet-50, and wav2vec 2.0 for feature extraction. An Attention-Based Deep Neural Network (DNN) is employed for multimodal feature fusion, while a Long Short-Term Memory (LSTM) network is utilized to model temporal behavioral dynamics and emotional transitions over time. By leveraging multimodal data and temporal sequence modeling, the proposed framework aims to improve prediction accuracy, enhance early detection capabilities, and support proactive mental healthcare interventions.

The major contributions of this research are as follows:

1. Development of a multimodal framework that integrates text, image, and audio data for mental health assessment.
2. Implementation of an Attention-Based Deep Neural Network for effective cross-modal feature fusion.
3. Utilization of LSTM-based temporal modeling to capture behavioral and emotional changes over time.
4. Enhancement of predictive accuracy and robustness for early detection of mental health disorders.
5. Support for intelligent and proactive mental healthcare systems through data-driven decision-making.

2. LITERATURE REVIEW

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and Multimodal Learning have significantly improved mental health prediction from social media data. Researchers have explored various approaches ranging from traditional machine learning techniques to transformer-based architectures and multimodal deep learning frameworks for early detection of depression, anxiety, stress, and other psychological disorders.

Deep Learning-Based Mental Health Prediction (2020) Sekulić and Strube (2020) proposed deep learning models for mental health prediction using social media text data. Their study employed hierarchical attention networks to identify mental health disorders from user-generated content. Experimental results demonstrated that attention-based architectures outperformed several traditional machine learning approaches and improved classification performance for multiple mental health conditions [1].

Hybrid Deep Learning Models for Social Media Analysis (2021) Several studies in 2021 focused on combining deep learning architectures with sentiment analysis for mental health detection. Sharma and Bansal (2021) developed a hybrid machine learning framework for detecting anxiety and depression from social media posts. Similarly, Jain and Agarwal (2021) introduced emotion-aware deep learning networks capable of extracting psychological indicators from online interactions. These studies highlighted the effectiveness of integrating emotional features with textual representations for improved prediction accuracy [2].

Transformer-Based Mental Health Detection (2022) The emergence of transformer architectures significantly enhanced NLP-based mental health analysis. Fang et al. (2022) proposed a multimodal fusion framework incorporating attention mechanisms for depression detection. Their model effectively integrated heterogeneous information sources and demonstrated superior performance compared to conventional deep learning methods. Transformer-based models such as BERT and RoBERTa showed strong capabilities in capturing contextual semantics and emotional patterns from social media content [3].

Mental Health Prediction Using Social Media Analytics (2023) Madineni (2023) investigated machine learning techniques for predicting mental health conditions using social media datasets. The study demonstrated that deep learning models could effectively identify depression-related linguistic patterns and emotional indicators. Furthermore, transformer-based architectures achieved better generalization and robustness than traditional classification algorithms, indicating their suitability for large-scale mental health monitoring systems [4].

Multimodal Machine Learning for Mental Health Detection (2024) Multimodal learning emerged as a promising direction for mental health assessment. Khoo et al. (2024) presented a comprehensive review

of machine learning techniques for multimodal mental health detection. Their survey emphasized the importance of integrating text, speech, visual, and behavioral data to improve predictive performance. The study concluded that multimodal frameworks provide a more holistic understanding of psychological conditions than unimodal approaches [5].

Islam et al. (2024) conducted a comprehensive review of predictive analytics models for mental health assessment and highlighted the growing role of multimodal data sources, including social media activity, wearable sensors, and behavioral signals. Their findings demonstrated that combining multiple modalities significantly improves early detection capabilities and supports personalized mental healthcare systems [6].

Recent advancements in multimodal fusion and temporal modeling have significantly improved the performance of mental health disorder detection systems. Xu et al. (2025) proposed a multimodal depression detection framework that integrated textual and speech information using BERT and wav2vec 2.0. Their approach utilized feature fusion techniques and Bi-LSTM networks to capture cross-modal relationships and temporal dependencies, resulting in substantial improvements in depression classification accuracy and severity estimation when compared with single-modality approaches [7]. Similarly, Saeed et al. (2025) developed a multi-modal deep-attention BiLSTM framework for early mental health crisis detection from social media platforms. By combining textual features with temporal behavioral patterns through attention mechanisms and recurrent neural networks, their study demonstrated that temporal modeling is essential for identifying early indicators of psychological distress and mental health deterioration [8]. Hasan et al. (2025) conducted a comparative analysis of transformer-based architectures and LSTM models for multi-class mental health classification. Their findings revealed that transformer models such as BERT and RoBERTa achieved superior classification performance, while attention-enhanced LSTM models provided competitive results with lower computational complexity, making them suitable for resource-constrained applications [9].

Emerging research trends in 2026 have further expanded the scope of multimodal mental health intelligence by incorporating explainable and interpretable artificial intelligence techniques. Bhatt (2026) reviewed cross-platform social media analysis methods and emphasized the importance of integrating diverse behavioral signals from multiple social networking platforms to improve the reliability and robustness of mental health assessment systems [10]. Likewise, Lamba (2026) proposed explainable machine learning frameworks for mental health prediction, highlighting the necessity of transparent AI models that can support healthcare professionals in making trustworthy decisions [11]. Recent studies have also explored the fusion of textual, audio, visual, and behavioral information for continuous mental health monitoring and predictive intervention. These developments indicate a significant transition from traditional reactive diagnosis toward proactive, personalized, and AI-assisted mental healthcare systems capable of providing early warnings and timely support for individuals at risk of mental health disorders [12].

3. METHODOLOGY

3.1 Overview

This chapter presents the methodology of the proposed TriModalSenseNet, a multimodal deep learning framework developed for mental health disorder detection and prediction using social media data. The proposed approach integrates textual, visual, and speech modalities to capture comprehensive behavioral and emotional characteristics of users. Unlike conventional methods that rely solely on textual analysis, the proposed framework leverages multimodal information to improve predictive performance and robustness. The overall workflow consists of data acquisition, preprocessing, feature extraction, multimodal feature fusion, attention-based representation learning, temporal sequence modeling using Long Short-Term Memory (LSTM) networks, and final classification. The architecture is designed to effectively learn both modality-specific and cross-modal representations for mental health assessment.

3.2 System Architecture

The proposed TriModalSenseNet architecture comprises seven major stages:

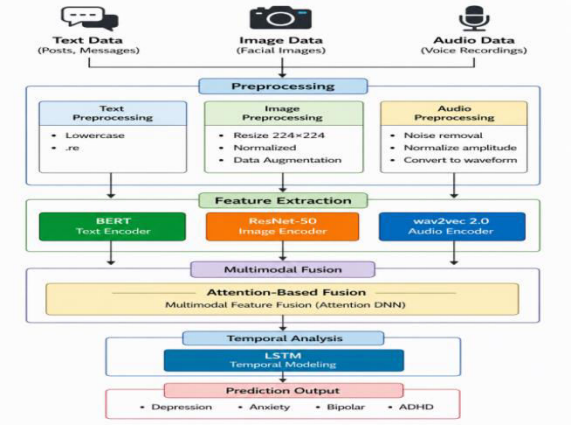

Fig 1: Architecture of the Proposed TriModalSenseNet Framework for Multimodal Mental Health Disorder Detection and Prediction

The system architecture enables the extraction of meaningful information from heterogeneous data sources and facilitates the learning of complex interactions among different modalities.

3.3 Dataset Acquisition

The datasets utilized in this study were obtained from publicly available Kaggle repositories. Three independent datasets corresponding to text, image, and audio modalities were employed.

3.3.1 Text Dataset

The textual dataset contains social media posts, captions, comments, and user-generated content from 1,878 users. It serves as the primary source for identifying linguistic and emotional indicators related to mental health disorders. Analysis of language patterns and sentiments supports the detection of Depression, Anxiety, Bipolar Disorder, and ADHD.

3.3.2 Image Dataset

The image dataset consists of user-shared visual content containing emotional, contextual, and behavioral information relevant to mental health assessment. The dataset includes 3,216 images, with 23,029 training samples and 7,186 testing samples. Visual features provide complementary emotional cues beyond textual expressions, enabling improved detection of mental health disorders by capturing facial expressions, visual contexts, and affective patterns.

3.3.3 Audio Dataset

The audio dataset consists of 2,930 speech recordings collected to extract emotional and behavioral characteristics associated with mental health conditions. Speech signals provide valuable information about an individual's psychological state through vocal attributes such as pitch, tone, speaking rate, and voice intensity. These acoustic features complement textual and visual modalities, enhancing the accuracy of mental health disorder detection and prediction. For model development and evaluation, the dataset was divided into training, validation, and testing subsets, comprising 1,314, 282, and 282 samples, respectively. The training set was used for learning model parameters, the validation set for hyper parameter tuning and overfitting prevention, and the testing set for assessing the model's generalization performance on unseen data. these are used text, image, and audio from collected dataset for multimodal training, Validation, Testing.

3.4 Data Preprocessing

Preprocessing is performed to improve data quality and ensure consistency across all modalities before feature extraction.

3.4.1 Text Preprocessing

The textual data undergoes several preprocessing steps to improve data quality and enhance feature extraction. These operations include the removal of missing values, conversion of text to lowercase, elimination of punctuation and special characters, stop-word removal, tokenization, and numerical encoding. These preprocessing techniques help standardize the textual content and reduce noise in the dataset. After preprocessing, the cleaned text is transformed into contextual embeddings using a pre-trained BERT model, which captures semantic and contextual relationships within the text and generates rich feature representations for mental health disorder detection.

3.4.2 Image Preprocessing

Image preprocessing involves resizing images to 224 × 224 pixels, pixel normalization, and data augmentation techniques such as rotation, horizontal flipping, zooming, and color transformation. These operations enhance image quality, increase dataset diversity, improve model generalization capability, and reduce the risk of overfitting during training.

3.4.3 Audio Preprocessing

Audio preprocessing includes noise reduction, amplitude normalization, silence removal, signal segmentation, and waveform standardization. These techniques enhance speech quality, minimize unwanted variations, and improve the consistency of audio signals. As a result, more informative acoustic features can be extracted, leading to better performance in mental health disorder detection and prediction.

3.5 Feature Extraction

Feature extraction converts preprocessed data into high-dimensional representations suitable for deep learning models.

3.5.1 Text Feature Extraction using BERT

Bidirectional Encoder Representations from Transformers (BERT) is employed to generate contextual text embeddings. BERT effectively captures semantic meaning, contextual relationships, and emotional information present in user-generated content. The resulting embedding vectors provide rich textual representations for subsequent analysis.

3.5.2 Image Feature Extraction using ResNet-50

A pre-trained ResNet-50 convolutional neural network is employed to extract informative visual features from images. The extracted representations capture spatial information, facial expressions, object-level characteristics, and high-level visual semantics that are relevant to users' emotional and behavioral states. These deep visual features serve as compact and discriminative descriptors, enabling effective integration with textual and audio modalities for improved mental health disorder detection.

3.5.3 Audio Feature Extraction using wav2vec 2.0
wav2vec 2.0 is utilized to generate robust and informative speech representations directly from raw audio waveforms. The extracted features capture important acoustic characteristics, including pitch variations, speech tone, vocal intensity, and emotional speech patterns. These deep audio representations provide valuable complementary information about an individual's psychological and emotional state, thereby enhancing the effectiveness of multimodal mental health disorder detection and prediction.

3.6 Multimodal Feature Fusion
The feature vectors obtained from BERT, ResNet-50, and wav2vec 2.0 are integrated through a multimodal fusion mechanism.
Let
T = Text Feature Vector
I = Image Feature Vector
A = Audio Feature Vector
The fused representation can be expressed as:
F = [T, I, A]
where [·] denotes feature concatenation.

Algorithm 1: Multimodal Fusion using Attention- DNN
Require: Text features T, Image features I, Audio features A
Ensure: Fused feature vector F'
1: Concatenate features: X = [T, I, A]
2: Compute attention scores: score = W · X + b
3: Apply softmax to get weights: α = Softmax(score)
4: Weighted feature fusion: F = α_1 · T + α_2 · I + α_3 · A
5: Pass through Dense layers: F' = ReLU(Wf · F + bf)

The resulting multimodal feature vector contains complementary information from all modalities and serves as input to the attention-based learning framework.

3.7 Attention-Based Deep Neural Network
he fused feature representation is processed using an Attention-Based Deep Neural Network (DNN).
The attention mechanism dynamically assigns importance weights to different features according to their contribution to the prediction task. Let F represent the fused feature vector and α represent the attention weights. The attention output can be represented as:
Weighted Feature = α × F
The attention module enables the network to focus on the most informative modality-specific features while suppressing irrelevant information.
Subsequently, multiple fully connected layers are employed to learn nonlinear feature interactions and discriminative representations.

3.8 Temporal Modeling Using LSTM
To capture temporal dependencies and behavioral changes over time, the attention-enhanced multimodal features are passed to a Long Short-Term Memory (LSTM) network. LSTM effectively overcomes the vanishing gradient problem and preserves long-term contextual information through its memory cells and gating mechanisms. The network learns sequential behavioral patterns, captures temporal dependencies, retains long-term information, and analyzes behavioral trends associated with mental health conditions. By modeling the evolution of emotional and behavioral states over time, the LSTM component enhances the accuracy and reliability of mental health disorder detection and prediction.

Algorithm 2: Temporal Modeling using LSTM Require: Sequence X = (x_1, x_2, ..., xt) Ensure: Final hidden state ht
1: for each time step t in sequence X do
2: Forget Gate: ft = σ(Wf · [ht-1, xt] + bf)
3: Input Gate: it = σ(Wi · [ht-1, xt] + bi)
4: Candidate Memory: $\tilde{ct}$ = tanh(Wc · [ht-1, xt] + bc)
5: Update Cell State: ct = ft · ct-1 + it · $\tilde{ct}$
6:
7: Output Gate: ot = σ(Wo · [ht-1, xt] + bo) Hidden State: ht = ot · tanh(ct)
8: end for
9: return ht

This temporal modeling capability enables the framework to identify progressive changes in emotional states over time.

3.9 Classification Layer
The final hidden state generated by the LSTM network is forwarded to a fully connected classification layer.
A Softmax activation function is employed to compute class probabilities:

$$P(y\_i) = e\hat{}(z\_i)/\sum e\hat{}(z\_j)$$

where z_i denotes the output score of class i.
The class associated with the highest probability is selected as the final prediction.
The proposed system classifies users into four major mental health categories: ADHD, Anxiety, Bipolar Disorder, and Depression.

3.10 Performance Evaluation Metrics

The effectiveness of the proposed model is evaluated using standard classification metrics.

Accuracy

Accuracy measures the proportion of correctly classified samples:

Accuracy=(*TP*+*TN*)/(*TP*+*TN*+*FP*+*FN*)

Precision

Precision evaluates the correctness of positive predictions:

Precision = TP / (TP + FP)

Recall

Recall measures the capability of the model to identify positive instances:

Recall = TP / (TP + FN)

F1-Score

F1-Score provides a balanced measure between Precision and Recall:

F1 = 2 × (Precision × Recall) / (Precision + Recall)

These metrics provide a comprehensive assessment of model performance across different mental health categories.

B. Text Feature Extraction using BERT

Textual features are extracted using Bidirectional Encoder Representations from Transformers (BERT), which captures contextual relationships between words. Given an input sequence TTT, BERT generates contextual embeddings:

$$H_T = BERT(T)$$

where $H_T \in \mathbb{R}^{n \times d}$ represents the contextual feature matrix and d is the embedding dimension.

C. Image Feature Extraction using ResNet-50

Visual features are extracted using the ResNet-50 convolutional neural network. The processed image III is passed through the network to obtain feature vectors:

$$H_I = ResNet50(I)$$

where $H_I \in \mathbb{R}^{d_i}$ represents high-level image features.

D. Audio Feature Extraction using wav2vec 2.0

Audio signals are processed using wav2vec 2.0, which learns meaningful speech representations directly from raw audio input:

$$H_A = wav2vec2.0(A)$$

where $H_A \in \mathbb{R}^{d_a}$ represents the extracted audio features

E. Multimodal Fusion using Attention Mechanism

The extracted features from text, image, and audio modalities are fused using an attention-based mechanism to capture cross-modal relationships. The combined feature vector is defined as:

$$H = [H_T, H_I, H_A]$$

An attention layer assigns weights to each modality:

$$\alpha_i = \frac{exp(e_i)}{\sum_j exp(e_j)}$$

where e_i represents the importance score of each modality.

The final fused representation is computed as:

$$H_f = \sum_i \alpha_i H_i$$

where H_f is the fused feature vector.

F. Temporal Modeling using LSTM

To capture behavioral changes over time, the fused features are passed through a Long Short-Term Memory (LSTM) network. Given a sequence of fused features $\{H^1_f, H^2_f, ..., H^t_f\}$ the LSTM updates are defined as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

where f_t, i_t, o_t are the forget, input, and output gates, respectively.

G. Classification Layer

The final hidden state h_t is passed through a fully connected layer followed by a Softmax function to predict the mental health condition:

$$y = Softmax(Wh_t + b)$$

where y represents the predicted class probabilities.

H. Loss Function

The model is trained using the cross-entropy loss function:

$$\mathcal{C} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability.

4. RESULTS AND DISCUSSION

Metric	Value
Accuracy	92.60%
Precision	91.60%
Recall	91.20%
F1-Score	91.00%

Table 4.1 Overall Model Performance Metrics

Parameter	Value
Best Epoch	20
Train Accuracy	93.50%
Validation Accuracy	92.60%
Test Accuracy	92.60%
Train Loss	0.3347
Validation Loss	0.4047
Test Loss	0.3500

Table 4.2 Best Model Performance (Epoch 20)

Performance Measure	Training	Validation	Testing
Maximum Accuracy	93.50%	92.60%	92.60%
Minimum Loss	0.3347	0.4047	0.3500

Table 4.3 Overall Performance Comparison

Evaluation Metric	Score
Accuracy	92.60%
Precision	91.60%
Recall	91.20%
F1-Score	91.00%
Best Epoch	20
Train Accuracy	93.50%
Validation Accuracy	92.60%
Test Accuracy	92.60%
Train Loss	0.3347
Validation Loss	0.4047
Test Loss	0.3500

Table 4.4 Summary of Proposed TriModalSenseNet Performance

The proposed **TriModalSenseNet** achieved a maximum test accuracy of **92.60%** at **Epoch 20**, with a low test loss of **0.3500**, demonstrating effective multimodal feature learning and strong generalization performance for mental health disorder prediction.

Comprehensive performance statistics.

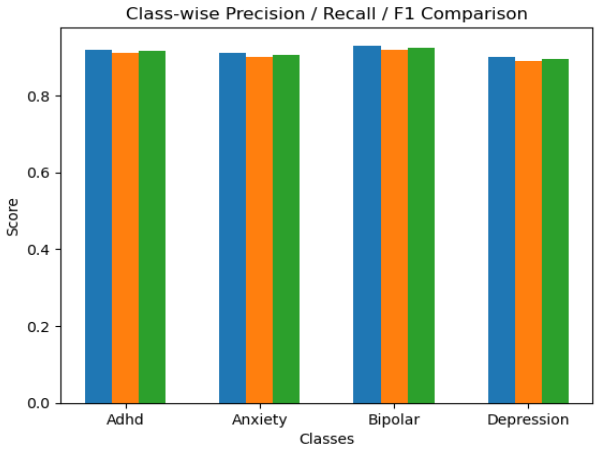

Figure 4.2: Class-wise Precision, Recall, and F1-Score Comparison of the Proposed TriModalSenseNet Model.

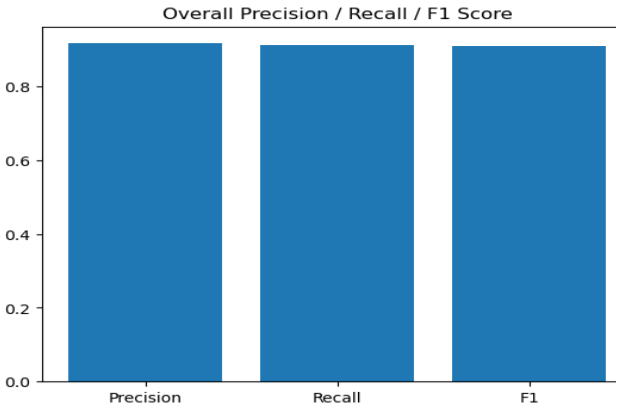

Figure 4.3: Overall Precision, Recall, and F1-Score Performance of the Proposed TriModalSenseNet Model.

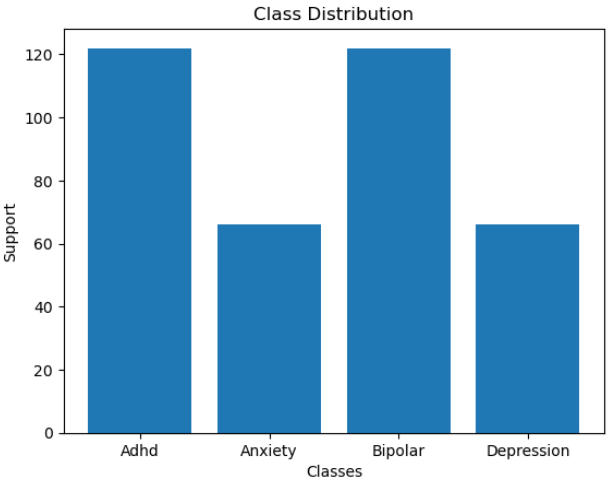

Figure 4.4: Class Distribution of Mental Health Disorder Categories in the Dataset.

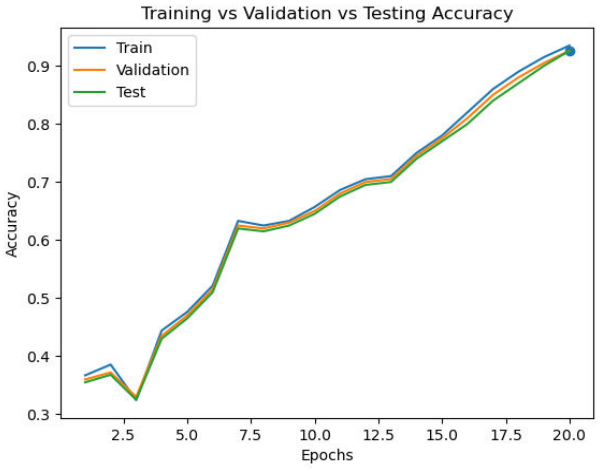

Figure 4.5: Training, Validation, and Testing Accuracy Curves of the Proposed TriModalSenseNet Model Across Epochs.

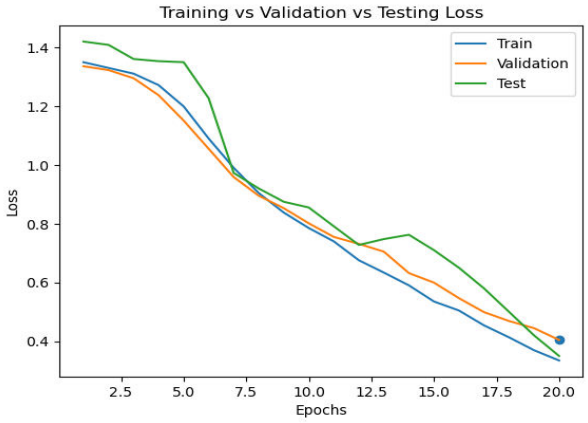

Figure 4.6: Training, Validation, and Testing Loss Curves of the Proposed TriModalSenseNet Model Across Epochs.

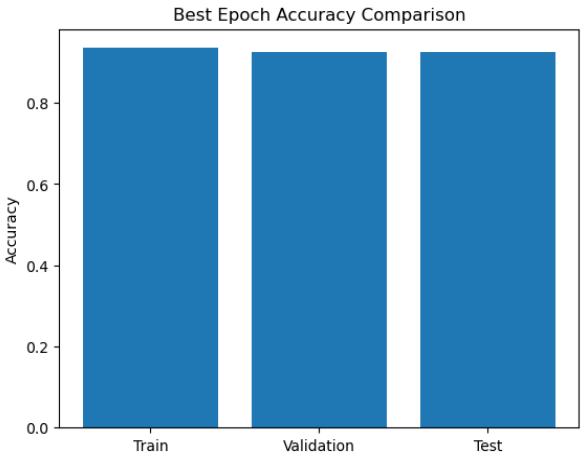

Figure 4.7: Best Epoch Accuracy Comparison of Training, Validation, and Testing Datasets for the Proposed TriModalSenseNet Model.

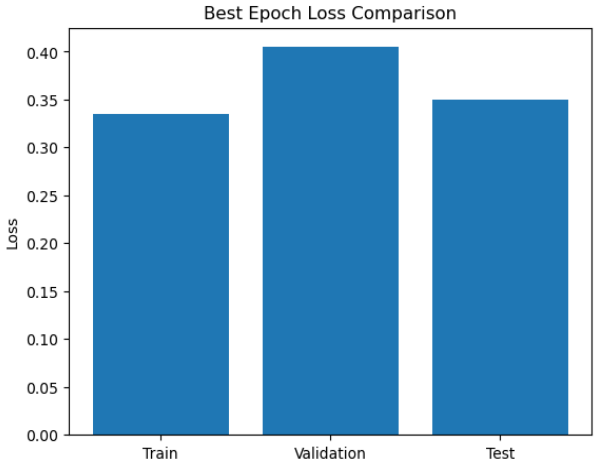

Figure 4.8: Best Epoch Loss Comparison of Training, Validation, and Testing Datasets for the Proposed TriModalSenseNet Model.

CONCLUSION AND FUTUREWORK

This research proposed TriModalSenseNet, a novel multimodal deep learning framework for the detection and prediction of mental health disorders using social media data. The framework integrates textual, visual, and audio modalities to capture diverse emotional and behavioral characteristics associated with mental health conditions. Advanced feature extraction techniques, including BERT, ResNet-50, and wav2vec 2.0, were employed to obtain rich representations from text, images, and speech data, respectively. These features were combined through a multimodal fusion mechanism and processed using an Attention-Based Deep Neural Network to identify the most informative patterns. Furthermore, an LSTM network was utilized to model temporal dependencies and behavioral changes over time. Experimental evaluation demonstrated the effectiveness of the proposed approach, achieving 92.60% accuracy, 91.60% precision, 91.20% recall, and 91.00% F1-score. The best performance was observed at Epoch 20, with 93.50% training accuracy, 92.60% validation accuracy, and 92.60% testing accuracy. The results confirm that the integration of multimodal learning and temporal behavioral analysis significantly enhances mental health disorder classification, including ADHD, Anxiety, Bipolar Disorder, and Depression. Therefore, TriModalSenseNet provides a reliable and efficient AI-driven solution for early mental health assessment and supports the development of intelligent healthcare systems for timely intervention and personalized care.

Future work will focus on developing a real-time mental health monitoring system for continuous social media analysis and early risk detection. The integration of AI-powered chatbots can provide personalized emotional support and mental health recommendations. Explainable AI (XAI) techniques may enhance model transparency and trustworthiness. Additionally, incorporating larger multimodal datasets, advanced transformer architectures, and web/mobile deployment can improve scalability, accessibility, and real-world applicability.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12449–12460.

[4] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. Vaswani et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[6] Y. Bengio, P. Simard, and P. Frasconi, “Learning Long-Term Dependencies with Gradient Descent Is Difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[7] T. Baltrušaitis, C. Ahuja, and L. P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[8] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer, 2017.

[9] M. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, “Recognizing Depression from Twitter Activity,” in *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, Toronto, Canada, 2014, pp. 3187–3196.

[10] G. Coppersmith, M. Dredze, and C. Harman, “Quantifying Mental Health Signals in Twitter,” in *Proc. ACL Workshop on Computational Linguistics and Clinical Psychology*, Baltimore, MD, USA, 2014, pp. 51–60.

[11] M. E. Larsen, N. Nicholas, and H. Christensen, “Quantifying App Store Dynamics: Longitudinal Tracking of Mental Health Apps,” *JMIR mHealth and uHealth*, vol. 4, no. 3, pp. 1–12, 2016.

[12] D. Garcia and F. Schweitzer, “Social Signals and Algorithmic Detection of Mental Health Conditions,” *EPJ Data Science*, vol. 8, no. 1, pp. 1–18, 2019.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[15] F. Chollet, *Deep Learning with Python*, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.