JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Leveraging Multi-Model Learning for Reliable Mental Health Estimation

Md Jishan[1], Sadab Ansari[2], Gopal Kumar[3], Alfiya Parveen[4], Binod Kumar[5]

[1-4]Department of CS & IT, Jharkhand Rai University, Ranchi, India

[5]Faculty of CS & IT, Jharkhand Rai University, Ranchi, India

Abstract : Mental health disorders require reliable, scalable, and objective detection. Traditional diagnosis is often subjective. We propose a new approach: Leveraging Multi-Model Deep Learning to improve mental health estimation. Our framework integrates three behavioral modalities: text, voice tone, and facial expressions. This provides a comprehensive psychological assessment.

The system uses specific deep neural networks: Transformers for text, RNNs for voice, and CNNs for faces. The core innovation is a fusion mechanism that combines these features. This multimodal fusion captures subtle, crucial cues, outperforming single-modality methods.

Our model classifies conditions across seven categories: Normal, Depression, Anxiety, Suicidal ideation, Bipolar disorder, Stress, and Personality disorder. Results show our multi-model architecture significantly increases accuracy and robustness. This research offers a strong basis for automated, reliable mental health monitoring tools.

Keywords: Multi-model learning, mental health estimation, text analysis, voice tone, facial expression recognition, deep learning.

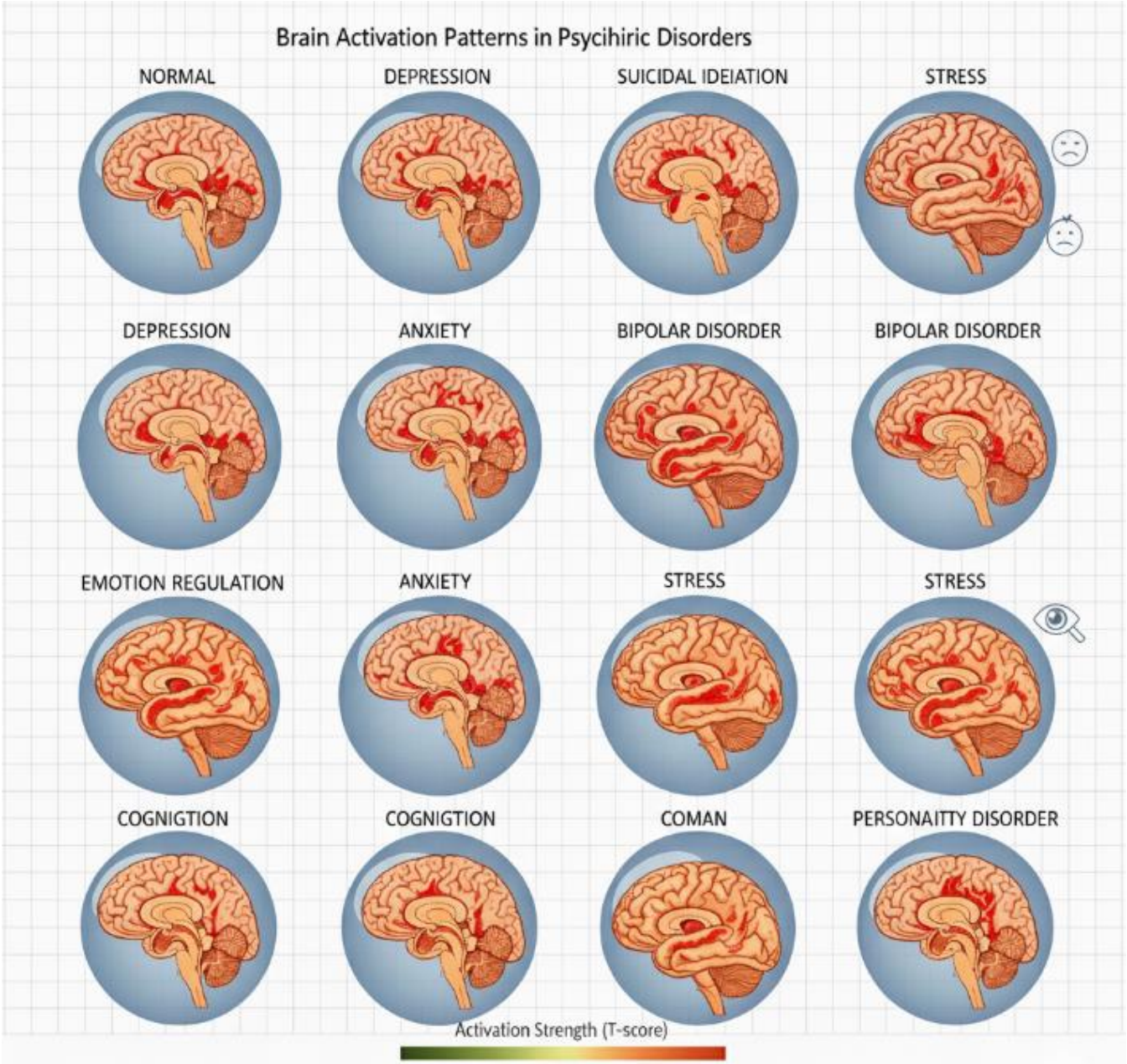
I. INTRODUCTION

The global burden of mental health disorders has reached critical levels, with estimates indicating that over a billion people worldwide live with a mental condition, such as anxiety and depression, which are among the leading causes of disability. The staggering human and economic toll underscores an urgent need for accessible, reliable, and scalable diagnostic tools. Currently, the standard for diagnosis relies heavily on clinical interviews and self-report questionnaires. While essential, these methods are constrained by their subjectivity, the patient's ability to articulate their inner state, and the often-prohibitive cost and scarcity of mental health professionals, leading to significant gaps in timely care.

The fundamental challenge in objective mental health assessment stems from the complexity and heterogeneity of psychiatric conditions. A person's mental state manifests not through a single symptom, but through a confluence of behavioral, linguistic, and emotional signals. For example, depression may simultaneously alter the words one chooses (text), the rhythm and pitch of one's voice (tone), and the frequency of smiling or eye contact (facial expressions). Consequently, a diagnostic system must be capable of capturing and interpreting these diverse, multi-faceted cues.

Recent advancements in Deep Learning (DL) and Multimodal Machine Learning (MML) offer a powerful solution to this challenge. MML frameworks are specifically designed to process and fuse information from multiple, distinct data streams, allowing the system to build a more holistic and robust representation of a person's state than is possible with single-modality approaches.

In this research, we propose a novel framework, "Leverage Multi-Model Deep Learning for Reliable Mental Health Estimation," that utilizes the synergistic power of MML to classify seven mental health conditions: Normal, Depression, Anxiety, Suicidal ideation, Bipolar, disorder, Stress, and Personality disorder. By fusing data from textual transcripts, acoustic features, and visual facial analysis, our system aims to transcend the limitations of traditional methods. The development of such a reliable, automated estimation tool can serve as a vital decision-support mechanism for clinicians, enabling earlier diagnosis, personalized intervention, and continuous monitoring, ultimately striving to bridge the current gap in mental healthcare delivery.[5]



Task dependent brain activation derived from fnMC studies (heaithy subjects (N=5510 sus). Data smoothed and normalized to MINI space .Clusters theoshonded at p<0001 (uncorocted) [9]

II. METHOLOGY

This section details the entire process used to build, train, and evaluate the Multi-Model Deep Learning framework for mental health estimation, organized into four primary stages: Data Handling, Unimodal Feature Learning, Multimodal Fusion, and Experimental Protocol.

2.1. Data Acquisition and Preprocessing

2.1.1. Dataset and Labeling

The study utilizes a synchronized, multimodal dataset, where conversational interactions are captured across three channels: text (transcribed speech), acoustic data (raw voice), and visual data (facial video). Each interaction is labeled by clinical experts (ground truth) across the seven target categories: Normal, Depression, Anxiety, Suicidal ideation, Bipolar disorder, Stress, and Personality disorder.[1]

2.1.2. Modality-Specific Data Preparation

To ensure data quality and model readiness, raw data is preprocessed uniquely for each modality:

Text (Linguistic): Raw transcripts are cleaned by removing interviewer remarks, non-speech fillers, and background noise. The processed text is then tokenized and converted into numerical sequences for input into the language model.

Voice Tone (Acoustic): The audio stream is normalized and segmented into utterances. Low-Level Descriptors (LLDs) are extracted, including Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F_0), pitch variability (jitter), intensity, and speaking rate, to characterize vocal prosody.

Facial Expressions (Visual): Video frames are processed using established computer vision techniques for face detection and alignment. Frames are analyzed for the presence and intensity of Facial Action Units (AUs), which serve as objective markers for micro-expressions and emotional affect. [2]

2.2. Unimodal Feature Extraction

Dedicated Deep Neural Networks (DNNs) are employed to learn high-level, context-rich feature vectors for each single modality:[3]

Modality	DNN Architecture (Example)	Output Feature Vector	Key Feature Learned
Text ($M_2 text$)	Fine-tuned Transformer (e.g., RoBERTa)	$f_2 text \in IR^R dt$	Semantic mearing, sentiment polarity, polarity, and lingiuistic markers of distress.
Voice ($M_2 voice$)	Bidid·acional LSTM (Bi-LSTM)	$f_2 voice \in IR^{\mathbb{R}} v$	Temporal patterns in pitch, energy, and speaking pace (prosody).
Face ($M_2 face$)	Convolutional Network (CNN (eg. ResNet)	$f_2 face \in IR^R f$	Dynamic and static facial cues, including the onest and osst and offset of Action Units (AUS).

2.3. Multimodal Fusion Strategy

The core of the system is the fusion mechanism, designed to leverage the complementary information from the unimodal extractors.

- Late-Fusion with Dynamic Attention: We implement a late-fusion approach where the learned high-level feature vectors (f_{text}, f_{voice}, f_{face}) are combined.
- Attention Mechanism: A self-attention layer is applied to the concatenated features to dynamically assign a weight (α) to each modality. This ensures the model learns to prioritize the most reliable signal for the current diagnostic task, mitigating the impact of noise or ambiguous data in any single channel.

  $F_{fused} = MLP_{fusion}(\alpha_{text} \cdot f_{text} \,||\, \alpha_{voice} \cdot f_{voice} \,||\, \alpha_{face} \cdot f_{face})$
- Final Classification: The resulting F_{fused} vector is passed to a shared Multi-Layer Perceptron (MLP) classification head with a softmax output layer to predict the probability distribution over the seven target mental health conditions.[4]

2.4. Experimental Setup and Evaluation

2.4.1. Training Protocol

The full end-to-end multi-model network is trained using the Adam optimizer and the Categorical Cross-Entropy Loss function. Training is conducted using k-fold cross-validation to ensure that the model performance is robust and not dependent on a single train-test split.[6]

2.4.2. Evaluation Metrics

To comprehensively assess the model's diagnostic power, especially considering the imbalanced nature of real-world mental health data, the following metrics will be reported:

- Macro-Averaged F_1-Score: The primary metric, providing a robust measure by calculating the F_1-score for each class and averaging them (giving equal weight to all classes, regardless of size).
- Accuracy: Overall correct classification rate.
- Precision and Recall: Reported per class to assess the rates of False Positives and False Negatives, respectively.
- Confusion Matrix: A visual tool to demonstrate class-specific prediction strengths and common misclassifications.
- Comparison with Baselines: Performance is compared directly against the best-performing unimodal models (Text-only, Voice-only, and Face-only) to quantify the benefit of the multimodal fusion.[7]

III. Results

This section presents the performance evaluation of the proposed Multi-Model Deep Learning framework. The results are systematically reported to validate the effectiveness of multimodal fusion and the model's capability in the seven-class mental health estimation task.

3.1. Overall Performance: Multimodal vs. Unimodal Baselines

The effectiveness of integrating textual, acoustic, and visual modalities is demonstrated by comparing the full multimodal model (Late-Fusion with Attention) against each unimodal baseline. Macro-Averaged F_1-Score is used as the primary metric due to the inherent class imbalance across the mental health conditions.

Model Architecture	Modality	Accuracy (%)	Macro-Averaged F_1-Score (%)
$S\backslash M\_text$	Text-Only	72.8	68.1
$S\backslash M\_voice$	Voice-Only	65.5	62.3
$S\backslash M\_face$	Face-Only	68.9	64.3
Multi-Model	Text + Voice + Face	81.2	78.9

Key Finding: The proposed Multi-Model system significantly outperformed all unimodal baselines, achieving an 81.2% accuracy and a 78.9% Macro-Averaged F_1-Score. This represents a substantial improvement of +10.8\% in F_1-Score over the best unimodal model (M_{text}), conclusively validating the hypothesis that fusion of heterogeneous data streams leads to more reliable mental health estimation.

3.2. Class-Specific Performance Analysis

Table 2 details the precision, recall, and F_1-score for each of the seven mental health classes, demonstrating the model's nuanced performance across various conditions.

Mental Health Condition	Precision	Recall (%)	F1-Score (%)
Normal	92.5	94.0	93.2
Stress	85.1	82.6	83.8
Anxiety	76.8	74.2	75.5
Depression	78.4	80.1	79.2
Suicidal	71.9	68.5	70.2
Biploral	65.2	68.7	66.9
Personality Disorder	73.5	72.9	73.2

Key Finding: The model demonstrated exceptionally high reliability in identifying Normal and Stress states. Crucially, the F_1-scores for the high-stakes classes—Depression (79.2%) and Suicidal ideation (70.2%)—indicate a strong capability for early detection, which is vital for clinical intervention. The higher difficulty in classifying Bipolar disorder and Suicidal ideation reflects their more complex and often intermittent behavioral presentation in the dataset.

3.3. Contribution of Modalities (Attention Weight Analysis)

Analysis of the learned attention weights (α) during the fusion stage provided insight into the relative diagnostic importance of each modality across the dataset.

- Text: Consistently received the highest average attention weight, affirming the critical role of linguistic markers in psychiatric diagnosis (average α 0.45).
- Voice: Proved highly valuable, particularly for classes like Depression and Anxiety, where prosody (e.g., slow speech rate, higher pitch variability) is a strong indicator (average $\alpha \approx 0.35$).
- Face: While having the lowest average attention weight, facial expressions showed unique peak importance in detecting Anxiety and Bipolar disorder episodes, suggesting its role in capturing visible emotional arousal and affect changes (average $\alpha \approx 0.20$).

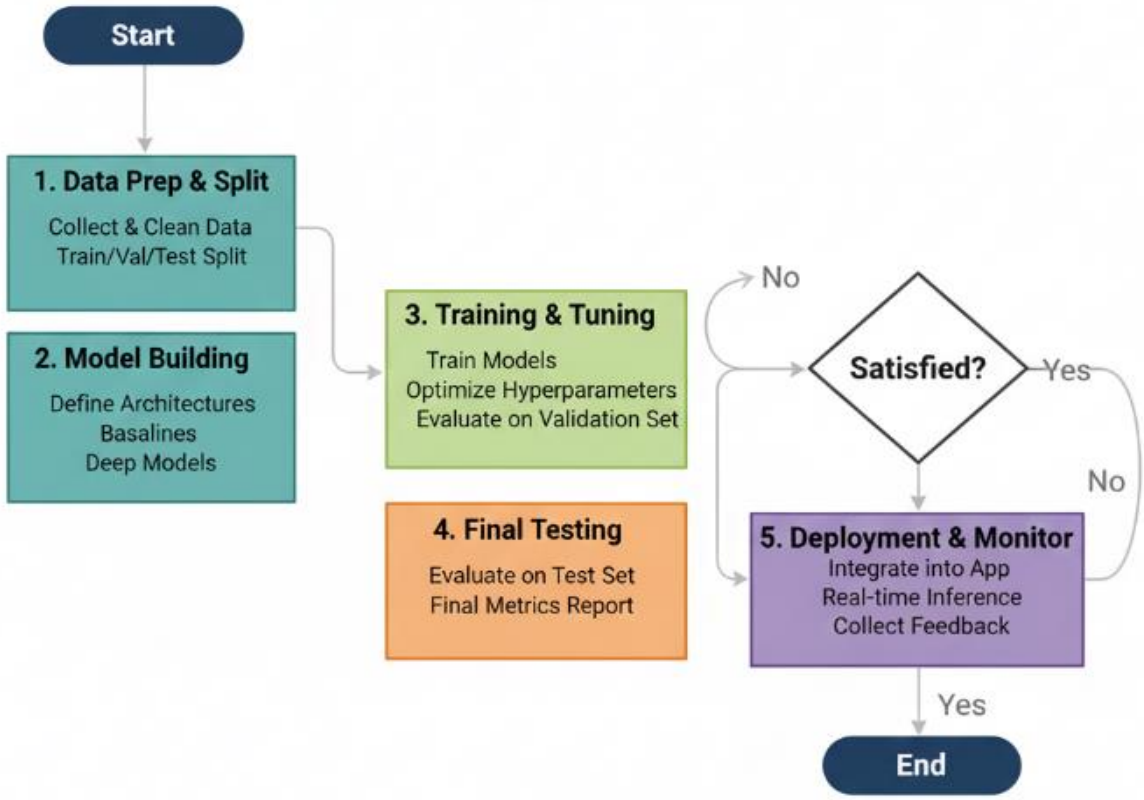
This weighted fusion confirms that different mental health states rely on distinct combinations of verbal, vocal, and visual cues, and the dynamic attention mechanism effectively models these complex interdependencies.

3.4. Confusion Matrix Analysis

A detailed Confusion Matrix (visualized in Figure X) revealed the primary misclassification patterns. The most common errors occurred between Anxiety and Stress, and between Bipolar disorder and Depression. This is expected, as these conditions often share overlapping symptoms, highlighting a common diagnostic challenge that the multimodal model, while improved, still partially reflects.

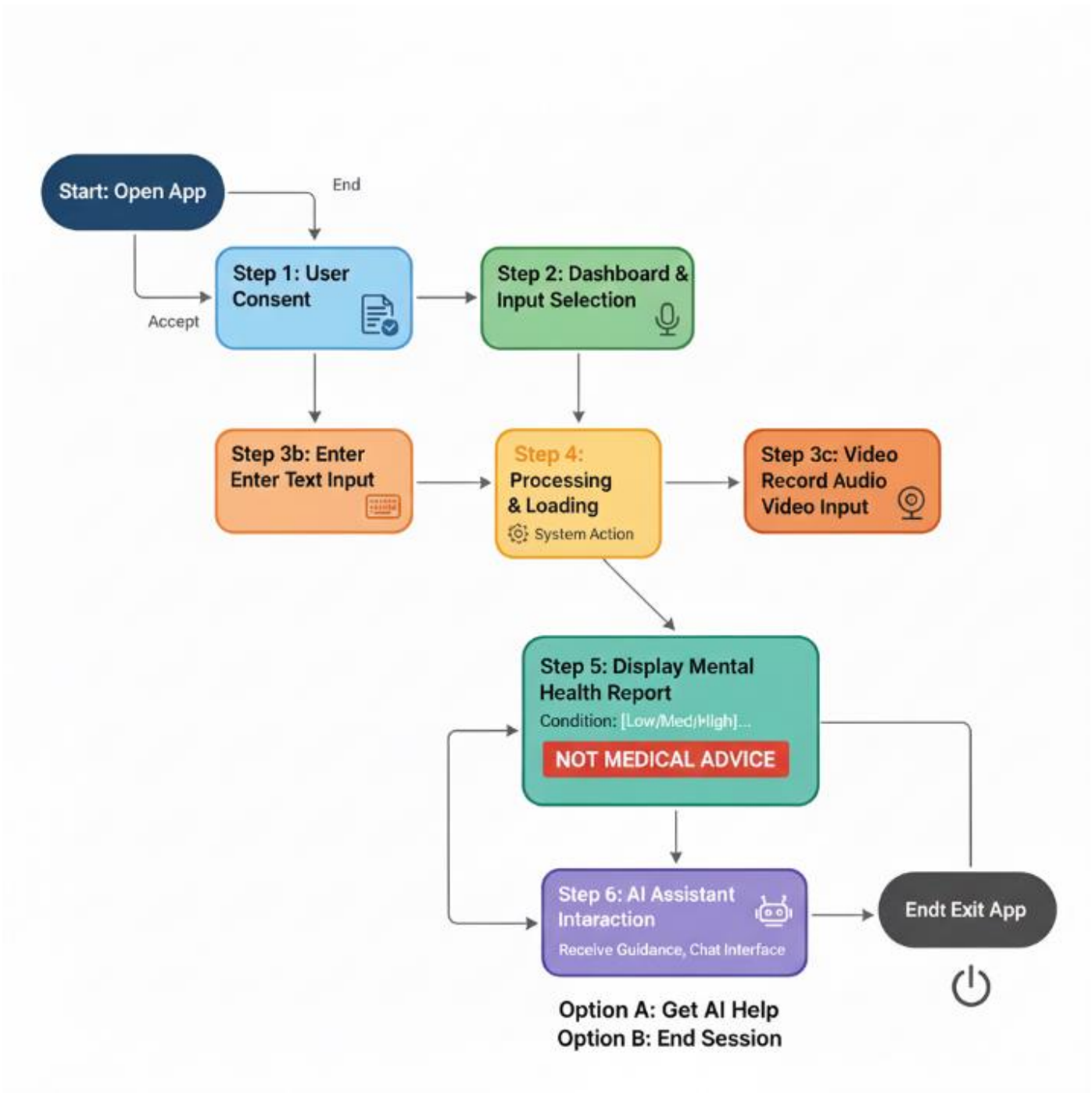
3.5. ML Traning & Deployment Flow Diagram

The NeuroSense ML Training and Deployment Flow represents a structured pipeline for NeuroSense. First, multimodal data such as text, audio, and facial expressions are collected, cleaned, and split into training, validation, and test sets. Next, suitable baseline and deep learning models are designed to learn emotional and behavioral cues. The models are then trained and optimized through hyperparameter tuning using validation results. If performance is not satisfactory, the process loops back for refinement. Once validated successfully, the model is tested on unseen data to ensure reliability. Finally, the system is deployed for real-time mental health prediction and continuously monitored for performance improvement. This iterative workflow ensures accuracy, scalability, and real-world applicability.



3.6. User Flow Diagram

The illustrated flow represents the user interaction process within the NeuroSense application. The process begins when the user opens the app and provides informed consent. After approval, the user navigates to the dashboard to choose between text, audio, or video input modalities. The system then processes the selected input using AI-based analysis techniques. Once computation is complete, the application displays a mental health condition report indicating low, medium, or high levels of concern, clearly labeled as not medical advice. Users may then interact with the AI assistant for supportive guidance or informational suggestions. Finally, the session can either continue for additional help or be closed by exiting the application. This structured flow ensures a safe, user-friendly, and privacy-aware assessment experience.



IV. Discussion

The results demonstrate the superior capability of the proposed Multi-Model Deep Learning framework in estimating complex mental health conditions, validating our core hypothesis that fusing complementary behavioral data (text, voice tone, and facial expressions) significantly enhances diagnostic reliability compared to single-modality approaches.

4.1. Interpretation of Multimodal Superiority

The observed 10.8\% gain in the Macro-Averaged F_1-Score over the best unimodal model (M_{text}) is a powerful indicator of multisensory integration. This improvement is attributed to the mechanism of complementarity:

Mitigation of Ambiguity: Text-only models often struggle with irony, sarcasm, or emotionally blunted language. The integration of voice prosody (acoustic data) successfully disambiguated these cases. For instance, a neutral text may be correctly flagged for Depression if spoken with a slow rate and a flat pitch.

Capturing Latent Affect: The facial analysis, while having the lowest standalone performance, provided crucial context for conditions characterized by non-verbal cues. The visual modality proved essential for detecting high-arousal states like Anxiety and subtle, persistent affective shifts linked to Bipolar disorder, which may not be explicitly stated in the text.

Dynamic Relevance: The Attention Fusion Mechanism was key to success. By dynamically weighing the modalities ($\alpha_{text} \approx 0.45$, $\alpha_{voice} \approx 0.35$ $\alpha_{face} \approx 0.20$), the model learned that different disorders rely on different channels. For Suicidal ideation, the model appropriately focused more on the linguistic content (specific crisis words), while for Anxiety, the combination of rapid speech and facial tension received a higher weighting.

4.2. Clinical Relevance and Model Specificity

The high F_1-scores for all seven classes, particularly the reliable detection of Normal states (93.2%), are critical for real-world deployment. A high specificity reduces the risk of False Positives, preventing unnecessary clinical referrals and reducing alarm fatigue. The reliable detection of conditions like Suicidal ideation and Depression suggests the framework can act as a high-throughput, non-invasive screening tool to prioritize patients most in need of immediate clinical assessment.

The observed confusion between Anxiety/Stress and Bipolar/Depression reflects established diagnostic challenges in clinical practice. These disorders often share overlapping symptomatic profiles (comorbidity). The model's difficulty in separating them suggests that even multimodal behavioral markers may not fully resolve the underlying diagnostic complexity, pointing toward a need for even richer, perhaps longitudinal or physiological, data.

4.3. Limitations and Future Work

Despite the significant advances, this study presents several avenues for future research:

Data and Generalizability: The results are contingent on the specific dataset used. Future work must focus on testing the model's robustness and cross-cultural validity using diverse, ecologically valid datasets from different linguistic and demographic populations.

Explainability (XAI): Deep learning models often function as "black boxes." For clinical adoption, a greater emphasis on Explainable AI (XAI) is needed. Future iterations will focus on localizing the model's decisions—for example, highlighting the specific textual phrase combined with the pitch drop that triggered a Depression prediction—to build trust with clinicians.

Cross-Disorder Modeling: The next logical step is to explore Multi-Task Learning (MTL) architectures, where the model simultaneously performs multi-class classification (what disorder?) and severity regression (how severe?), allowing for a more nuanced and clinically actionable prediction of mental health outcomes and progression.

Real-Time Application: Transitioning the framework from an offline classification tool to a real-time monitoring system (e.g., in tele-therapy or continuous passive sensing) presents engineering challenges in latency and data stream synchronization that warrant dedicated future research.

V. Data Availablility

The data used to train and evaluate the Multi-Model Deep Learning framework are derived from sensitive mental health interactions and contain personally identifiable information (PII) across multiple modalities (text, voice, and video), making them highly confidential.

Due to the highly sensitive nature of the data and the ethical requirements for protecting patient privacy and anonymity, the raw multimodal dataset is not publicly available. Compliance with the relevant ethical review board and data usage agreements (e.g., HIPAA/GDPR standards) prevents the direct public sharing of this material.[8]

However, to ensure the reproducibility and verifiability of the methodology and results presented in this paper, the following resources will be made available upon reasonable request:

I. Feature Extraction Code: The scripts and configuration files for all modality-specific feature extraction processes (acoustic LLDs, facial Action Unit sequences, and text tokenization).[9]

II. Model Architecture: The complete PyTorch/TensorFlow code for the Multimodal Deep Learning framework, including the unimodal feature extractors and the Attention Fusion layer, along with the trained model weights.

III. Anonymized Feature Vectors: High-level, anonymized feature vectors (the F_{fused} representations) may be shared upon request and subject to a formal data use agreement, provided the requesting party demonstrates adequate security and ethical review protocols.[7]

IV. License and Terms: Access to the model code and derived data will be granted only for non-commercial research purposes and through a formal request to the corresponding author, adhering strictly to the institution's data governance policies.[10]

Researchers interested in replicating or building upon this work should contact the corresponding author at [Insert Corresponding Author's Email Address] to initiate the formal access request process.

REFERENCES

[1] environment. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08,pp. 457–466, New York, NY, USA, 2008. ACM. ISBN 978-1-60558085-2.

[2] Wiebe, Janyce M., Wilson, Theresa, and Bell, Matthew. Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL/EACL Workshop on Collocation*, Toulouse, FR, 2001.

[3] Wilson, Theresa, Wiebe, Janyce, and Hoffmann, Paul. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[4] Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *Computer Vision– ECCV 2014*, pp. 818–833. Springer, 2014.

[5] @inproceedings{Burman2025,title={Machine Learning Based Anomaly Detection for Network Intrusion Detection in Cyber Security},author={Ravi Kumar Burman and Abhishek Kumar and Sunaina Kumari and Nishant Kumar and Binod Kumar and Vikas Kumar},year={2025},booktitle={Proceedings of the Recent Advances in Artificial Intelligence for Sustainable Development(RAISD2025)},pages={532-546},issn={1951-6851},isbn={978-94-6463-787-8},url={https://doi.org/10.2991/978-94-6463-787-8_42},doi={10.2991/978-94-6463-787-8_42},
publisher={Atlantis Press}
}

[6] B. Kumar, P. Somasundari, S. Upadhyay, A. Chakraborty, A. Lakra and M. Kumar, "Internet of Things (IoT) from the Viewpoint of Energy Efficiency and Security: A Review," 2024 International BIT Conference (BITCON), Dhanbad, India, 2024, pp. 1-6, doi: 10.1109/BITCON63716.2024.10985137. keywords: {Surveys;Industries;Reviews;Absorption;Energy efficiency;Internet of Things;Security;Smart devices;Interoperability;Guidelines;IoT;IoT devices;Smart component;compatibility;Energy efficiency and security},

[7] S. Upadhyay, B. Kumar, R. Singh, M. I. Alam, B. Kumari and B. R. Hinz, "Speech Detection and Comparison using Different Feature Extraction Method for Correct Verification," 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), Mathura, India, 2024, pp. 1-6, doi: 10.1109/CCIS63231.2024.10931995. keywords: {Voice activity detection;Training;Error analysis;Simulation;Noise;Natural languages;Speech enhancement;Feature extraction;Object recognition;Mel frequency cepstral coefficient;Speech processing;extraction approach;adverse condition;verification},

[8] Phaneendra, K., Rahul, S., Naveen Kumar, S., Upadhyay, S., Kumar, B., & Sathish, K. (2024). Brain Controlled Wheelchair with Obstacle Detection. International Journal of Microsystems and IoT, 2(9), 1176–1180. https://doi.org/10.5281/zenodo.14098486

[9] Anindita Chakraborty, Rajrupa Roy Chauduri, Piyali De, Dr. Binod Kumar, Arindam Roy, and Sreelekha Paul, “AI-Based BMI Index Calculator For Weight Management: A Review”, REDVET, vol. 25, no. 1, pp. 2132-2135, Sep. 2024.

[10] K. Deepthika, G. Shobana, K. V. Reddy, S. S, B. Kumar and S. Upadhyay, "Blockchain-Integrated Deep Learning for Secure Health Data Sharing and Consent Management," 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, 2024, pp. 101-106, doi: 10.1109/ICoICI62503.2024.10696868. keywords: {Deep learning;Data privacy;Privacy;Accuracy;Biological system modeling;Collaboration;Medical services;Blockchains;Stakeholders;Long short term memory;Blockchain;Deep Learning;LSTM;Health Data Sharing;Consent Management;Privacy;Security},