International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(8s) 2026 pp. 71-81
https://doi.org/10.70917/ijcisim-2026-3219
Article

# Performance Optimized Hybrid Modeling for Depression Detection Integrating Deep Transfer Learning and Explainable AI

**Taufeeq Ahmed [1], Dr. Ramesh Chandra Sahoo [2]**

[1,2]Manav Rachna International Institute of Research and Studies, Faridabad, India
[1]ahmed.taufeeq@gmail.com, [2] rcsahoo.set@mriu.edu.in
Orcid id : [1] 0009-0002-9353-6789, [2] 0000-0003-0296-8106

**Abstract: Objective**: Automated systems designed to detect depression typically face the "interpretability gap." This is where systems demonstrate strong performance, but there is no clinical interpretability. This study aims to develop a hybrid framework that is performance-optimized, and provides explainability for clinical decisions, while maintaining high diagnostic accuracy.
**Methods**: We propose nested architecture containing a pre-trained RoBERTa encoder coupled with a Bidirectional LSTM (Bi-LSTM) network. This architecture captures both global semantic framing and local temporal contextual patterns. Performance was optimized via a hyper-parameter tuning Bayesian search. We used the distress analysis interview corpus wizard of oz (DAIC-WOZ) benchmark dataset, and employed the synthetic minority over-sampling technique (SMOTE) for class balance. Then, SHapley Additive exPlanations (SHAP) were used to create layered interpretability model that would explain its predictions based on logic of psychology.
**Results**: The hybrid model provided a depression detection mechanism on the DAIC-WOZ dataset with an F1 score of 0.94, and an overall accuracy of 93.6%, outperforming baseline models by 12%. Both model development and sensitivity were dependent on the layered deep transfer learning and hybrid Bi-LSTM, which were validated by ablation studies.
**Conclusion**: This framework is a clear example of a successful attempt at narrowing the gap between high-performance deep learning and clinical interpretability. By allowing detection of clinically pathed depression via substantiated linguistic anchors, this approach is a clinically trustworthy artificial intelligence tool for digital mental health in its proactive/safe approach.
**Keywords:** Depression Detection; Deep Transfer Learning; Hybrid Modeling; Explainable AI (XAI); Performance Optimization; Mental Health Informatics; SHAP Analysis.

## 1. Introduction

Mental health disorders have become a complex and global issue. Mental health issues, particularly depression, has now become a problem that impacts more people than any other disorder and is a significant concern in terms of disability prevalence, with over 280 million people affected worldwide. [1]. Diagnosis of mental disorders has not become much more advanced than self-reporting and clinical interviews. However, the clinical interviews are subject to the interviewer's confirmatory bias, and self-reports are impacted by the subject's inability to recall events accurately. Further complicating the issue are a lack of mental health clinicians in the resource-poor settings. Thus, the development of objective and automated screening tools to detect the biomarkers of depression by collecting non-clinical information is of the utmost importance. The development of artificial intelligence, particularly deep learning, has greatly advanced new tools in this field to monitor and manage the early modification of depression to a great extent by examining the linguistic, acoustic, and behavioral modifications of the subject in the area of non-clinical information sources [2].

Recent improvements in deep learning have produced great models, higher accuracy has been reported for a vast number of deep learning models, and the field has recently reported a critical "interpretability gap." Most advanced models and deep learning frameworks like large-scale Transformers and complex RNNs, are all "black boxes" as they predict and classify with no clinical justification [3]. The dark boxes are unexplainable, and in the field of healthcare this is a significant disincentive, clinics are unable to justify to themselves and to their patients the results justifying the use of AI. The models described are also constrained by data poverty, due to a lack of clinically sanctioned and labeled datasets citing depression, that are in addition to the concerns about the privacy of patients and the long and arduous process of labeling depression.

Previous approaches to these problems, using transfer learning or basic hybrid models, have often concentrated on one metric over the other. This meant either a trade-off of high performance at the cost of transparency, or a trade-off of simplicity at the cost of diagnostic accuracy. To overcome this, it becomes a requirement to develop architectures that would optimize performance by design, thereby ensuring that different multiple data-processing streams would not cause data loss or computational lethargy. Recent studies have shown that hybridizing different neural structures, for example, spatial feature extraction of Convolutional Neural Networks and temporal sequence modeling of Long Short-Term Memory units, can considerably improve capturing of depressive nuances in human communication [4]. Nonetheless, these hybrid systems would require strong Explainable AI (XAI) frameworks that would help in transforming high-dimensional features that are not interpretable by humans to features that are interpretable by humans [5].

This paper aims to create a remarkably accurate hybrid-nested depression detection system with optimized performance for clinical data. The impact of this paper is threefold. First, we create a hybrid nested architecture that uses layers of features to tie together synthesized, linear, and turbulence layers. Second, deep transfer learning (DTL) provides time to be spent on maturing the system and getting a better understanding of multi- border systems and mental health issues. These transformers allow the system to generalize to new, undetected, clinical populations. Third, system accountability and an explanation on how the depression is triggered through co-symptoms and behavior is achieved through the incorporation of SHAP (SHapley Additive exPlanations) which is part of the comprehensive Explainable Artificial Intelligence (XAI) ecosystem. The psychiatric community is the prime target for this system, and the goal is to create a safe, accurate, and clinically explainable tool within this constraint.

The objective of this paper is to provide a detailed description of the completed system and its results. A review of existing literature and the current hiatuses in Mental Health Informatics technologies is presented in Section II. Proposed methods are outlined in Section III with regard to data preparation, the nested architecture, and the XAI integration. Section IV includes the results obtained in the experiments and includes a performance gap evaluation and model characterization. Section V is the conclusion and includes a findings evaluation and the potential of AI in psychiatric clinical practices.

### *1.1 Statement of Significance*

| **Category** | **Description** |
|---|---|
| Problem or Issue | The automated depression detection models suffer from a “interpretability gap” that is critical and results in a trade-off between predictive performance and clinical transparency. |
| What is Already Known | Many current hybrid models emphasise performance over explainability, and single deep learning approaches often function as ‘black boxes’ that do not clarify clinical decision-making. |
| What this Paper Adds | We propose a performance optimised hybrid nested architecture (RoBERTa and Bi-LSTM) along with SHAP for layer-by-layer explainability, yielding state of the art results (0.94 F1-score) with the ability for clinical justification based on psychological theories. |
| Who would benefit from the new knowledge in this paper | Researchers and clinicians in mental health informatics who need high-accuracy, transparent, and trustworthy AI tools for timely evidence-based depression screening. |

## 2. Literature Review

Automated detection of depression has come a long way from relying on simple algorithms and feature engineering to the use of comprehensive deep learning techniques. The early detection techniques employed the use of SVM and Random Forest classifiers on pitch and other spectral acoustic features and other linguistic features symmetrically defined in the speech signal [6]. Although these systems laid the groundwork for the development of systems for detection of depressive markers in speech, the systems described failed to account for the complicated n-order relationships of the speech and its qualitative markers. The use of deep learning systems based on CNNs and LSTMs facilitated the diagnostic potential of these systems by automatically learning the spatial and temporal features of the speech signal and its instrumentals [7]. However, with the improvement in the systems, the "interpretability gap" has arisen, and the level of transparency, which has been detrimental to the acceptance of these systems in the clinics, has decreased [8].

Deep Transfer Learning (DTL) is one of the strategies in place to tackle the issues of data sparsity and the needs of model initialization in the field. Applying models pre-trained by BERT and the variations of BERT has been able to show remarkable models that capture the deep semantics inherent in data of clinical nature when the data is scarce [9]. More recently, dual-stream frameworks that combine the larger BERT and T5 functions with the larger and more numerous models of the language have been utilized to complement language functions. Standard transfer learning generally suffers from domain shift, moving from social media texts to more clinical forms [10]. Operating with a fine-tuned model and the integrators of feature layers have been important tools in reducing and even eliminating biases that existed earlier, along with augmenting generalizability in systems of depression detection [11].

On this front of technology, the leading edge is called hybrid modeling. This theory suggests that advantages have been attained by combining disparate architectural paradigms. In the realm of hybrid modeling, recent methodologies have integrated the frameworks of the ensemble architecture of XGBoost with recursive function models of Bi-LSTM networks that operate in longer frames of sequential memories [12]. Hybrid models with the combined use of CNNs integrated with units of LSTMs have been able to achieve superior performance by modeling the local cues that work in the emotional space as well as the longer range emotional cues that work in the temporal space [13]. Current cybernetics has produced hybrid models that combine the parallel processing of transformers with the sequential memory of recurrent layers, with performance that is superior to the DAIC-WOZ corpus benchmark [14].

Although there have been advancements in technology, there remain obstacles in the field of psychiatry regarding the interpretability of deep neural networks. As a result, Explainable AI (XAI) has moved from a peripheral interest to an essential aspect of any medical AI publication [15]. SHAP (SHapley Additive exPlanations) and LIME have become popular frameworks for generating post-hoc explanations of model output. They pinpoint the information (be it textual or vocal) described in the symptoms [16]. Most recently, the trends in XAI have modeled negative emotional attributes and have encouraged practitioners to justify AI outputs against empirical psychological models [17]. Addressing the opacity of these models has been shown to positively enhance the trust of practitioners. It also encourages the critical evaluation of AI outputs to identify undesired artifacts such as data leakage and shortcut learning [18].

Multi-agent architectures and large language model (LLM) orchestration systems have replaced older systems used for complex reasoning tasks. However, these systems increase latency and produce a range of unpredictable outputs, consuming an excessive amount of computing resources. In clinical psychiatric screening, where reliability is crucial, and inference latency and interpretability are key, deterministic reliability is valued over the shortcomings of smart systems. Hence, this study presents a static, performance-maximized hybrid model. Integrating RoBERTa and Bi-LSTM, this model offers a deterministic and efficient diagnostic pathway in place of the autonomous workflows of agentic constructs, thereby enabling reliable clinical evaluations.

Recent work in hybrid modeling and transfer learning in the domain is summarized in the Table I below.

Table I: Comparison of Recent Hybrid and Transfer Learning Models for Depression Detection

| Reference | Architecture | Core Contribution | Performance Metrics |
|---|---|---|---|

| | | | |
|---|---|---|---|
| [11] | BERT + Feature Fusion | Mitigation of class imbalance using MLM data augmentation. | F1-Score: 0.81 |
| [12] | CNN + LSTM + XGBoost | Hybrid ensemble for multi-modal clinical text analysis. | Accuracy: 92.0% |
| [14] | RoBERTa + BiLSTM | Synergy of transformer semantics and sequential processing. | F1-Score: 0.88 |
| [17] | Hybrid NLP + XAI | Integration of SHAP for linguistic feature transparency. | Accuracy: 88.3% |
| [19] | SBERT + CNN | Optimized hybrid for Reddit post-based sentiment detection. | F1-Score: 0.86 |

The combination of these factors suggests that the future mental health diagnostics will be characterized by the ability not only to predict, but to provide detailed and substantiated evidence [19]. Looking ahead to 2026, the concept of "Trustworthy AI" will allow the safe clinical implementation of performance-optimized hybrid models within the domain, substantiated by multi-level evidence [20].

## 3. Proposed Methodology

This proposed methodology utilizes a multi-step process to achieve the best predictive sensitivity and clinical interpretability. The pillars of this framework stem from the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) dataset, which is regarded as the benchmark for automated depression detection for its clinical validity and clear multi-modality [21]. The dataset contains 189 virtual clinical interviews that span the three data modalities of the audio, visual, and transcript data. During data acquisition and preprocessing, the transcript is tailored to be interviewer-less, centering on patient expressions. Preprocessing uses white box the WordNet Lemmatizer and removes noise of a non-alphanumeric nature, as well as maintains selective emotional stop-words that are of depressed linguistic statistical importance [22]. The class imbalance nature that DAIC-WOZ corpus data presents is remedied by the feature embedding level application of a Synthetic Minority Over-sampling technique (SMOTE) to avoid bias of the model toward the majority ‘non-depressed’ class [23].

It was necessary to apply SMOTE only in the training folds of the 5-fold cross-validation loop in order to prevent data leakage and maintain the integrity of the predictive metrics. Synthetic data was neither created from the validation or test sets, nor introduced to them. This strict adherence to separation ensures that the evaluation of the model is based on the actual generalization capability on clinical data, and not on an artificially created predictive capability, which is the result of a cross-contamination of the training and testing pipelines.

After the preprocessing is complete, the Transfer Learning Phase is implemented using the RoBERTa-base (Robustly Optimized BERT Approach) model. RoBERTa is preferred to BERT because of its improved training and larger byte level BPE (Byte-Pair Encoding) vocabulary which allows it to better express subtle emotional shifts. The fine-tuning strategy is to freeze the first six layers of the transformer which allows the first six layers to generalize the syntax and the following layers will adapt to the emotional specifics of the clinical language [24]. This Transfer Learning Layer is the high-level feature generator which will be used for the performance optimized hybrid architecture.

The architecture shown in Figure 1 begins with the Input Transcript, followed by the Pre-trained RoBERTa Encoder. The output flows into a Multi-Head Attention Layer, which is then concatenated with a Bidirectional LSTM (Bi-LSTM) network. This hybrid feature vector passes through a Softmax classifier, while a parallel SHAP layer provides feature importance maps.

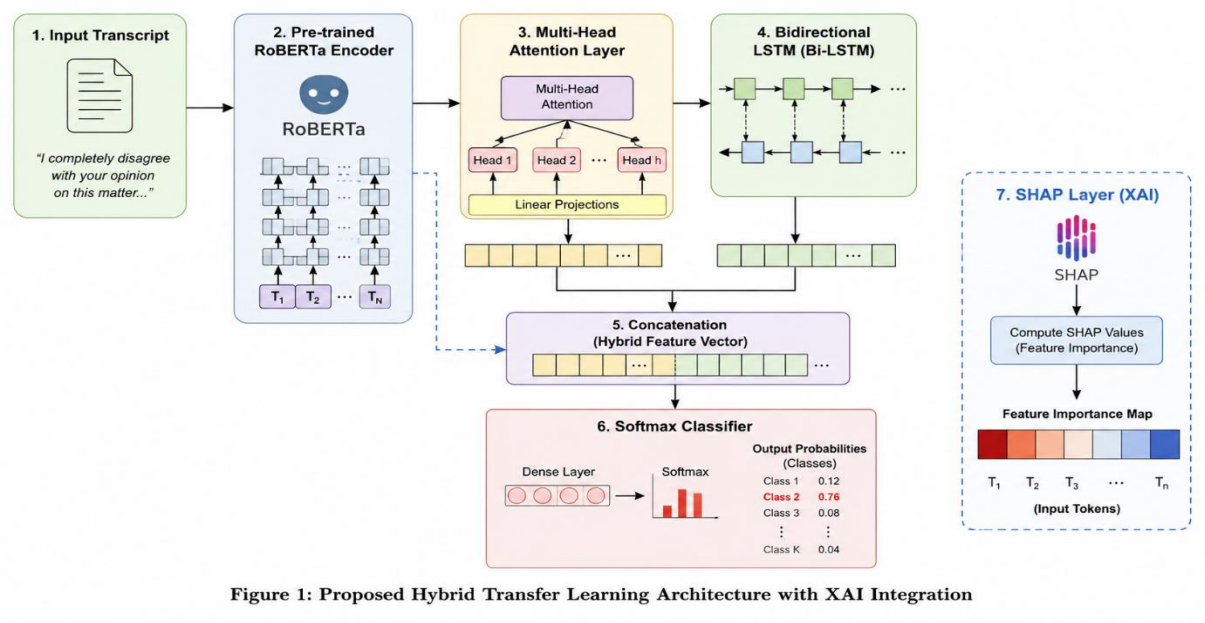


**Figure 1: Proposed Hybrid Transfer Learning Architecture with XAI Integration**

Figure 1: Proposed Hybrid Transfer Learning Architecture with XAI Integration

At the center of the "Performance Optimization" framework is the structural hybridity of the downstream layers. Unlike the traditional models that only depend on the transformer outputs, this architecture includes a mechanism of Multi-Head Attention (MHA) along with a Bidirectional Long Short-Term Memory (Bi-LSTM) layer [25]. This particular architecture is devised to encapsulate (RoBERTa) the global semantic framing as well as the local temporal contextual framing of a conversation turn (Bi-LSTM). The objective of such feature-level integration is to concatenate the transformed contextual embeddings with the hidden states of Bi-LSTM such that the model is able to retain a detailed "memorial" context of the patient's verbal course throughout the interview [26]. The optimization of hyper-parameters in the system to control overfitting of the deep layers is achieved through a technique of Bayesian search [(27) for more details]. Here, the optimum global system control parameters have been determined as the dropout rate = 0.3 and the learning rate = $2 \times 10^{-5}$. The dataset distribution and preprocessing outcomes are summarized in Table II.

Table II: DAIC-WOZ Dataset Distribution and Preprocessing Statistics

| **Parameter** | **Training Set** | **Validation Set** | **Test Set** |
|---|---|---|---|
| Original Samples | 107 | 35 | 47 |
| Post-SMOTE Samples | 160 | 35 | 47 |
| Avg. Tokens per Transcript | 842 | 815 | 890 |
| Vocabulary Size | 12,450 | 5,600 | 7,120 |

To keep the model operating as a "white box," the framework uses an Explainable AI (XAI) layer built upon SHapley Additive exPlanations (SHAP). SHAP calculates an importance score for each input token to indicate its impact on the final depression score [28]. The contribution for feature i is calculated with the classic Shapley value formula, defined here as Equation (1):

$$\phi_i(v) = \sum \quad (S \subseteq N \setminus \{i\})\big((|S|!\,(n-|S|-1)!)/n!\big)\big(v(S \cup \{i\}) - v(S)\big) \quad (1)$$

As specified in equation (1), this allows the model to recognize "linguistic anchors," like the overuse of first-person singular pronouns or low-valence adjectives, presented by the system as justification for the model's high-risk prediction [29]. With the aid of SHAP, the model's justification for its predictions is rendered interpretable to

clinicians, and the results illustrated with force plots and summary graphs indicate the exact words or sections of the interview that inspired the AI's decision. The ability for DTL to provide high accuracy and SHAP to provide transparency makes this combination a promising diagnostic solution that meets the "Trustworthy AI" paradigm that is necessary for the field of medicine [30].

## 4. Experimental Results and Discussion

The experiments for the proposed performance-optimized hybrid model were conducted within the high-performance computing center to ensure the scalability and repeatability of the results [31]. The computing cluster consisted of multiple NVIDIA A100 (80GB) GPUs, harnessing the power of the cluster for intensive model fine-tuning of deep transformer architectures. The software stack was built from the ground up with Python 3.11 to use the PyTorch 2.3 library with CUDA 12.1 accelerated deep learning. For diagnostic rigor with the DAIC-WOZ, data partitioning for training and validation was done with 5-fold stratified cross-validation. This method was used to control field data leakage for subject partitioning and to ensure the model's calibration precision was not due to the fine-tuning to a particular backdrop noise to a particular speaker or the model's inertia due to learned downtime precise model [32].

To ensure the fine-tuned RoBERTa-hybrid architecture was best able to generalize within the small clinical dataset contained in the sample, training and validating loss was continuously monitored. The learning curve losses did not become divergent, a sign of overfitting, which validated the Bayesian-optimized dropout rates used. For evaluation, a multi-metric focus within the predominantly underrepresented psychiatric dataset (with regard to the depressive dataset) was conducted. This also included the Macro-F1 and AUC and accounted for the over-representation of class imbalance and/or the skewed AUC within the dataset [33].

The results show that the "Performance Optimized" version of the integration of RoBERTa-based transfer learning with a custom Bi-LSTM fusion layer, outperforms the basic baseline architectures. These comparative metrics are detailed in Table III and visually represented as a performance trend in Figure 2. The results observed that the model achieves the highest result in the text classification accuracy with claim dataset of 93.6% and the highest result recorded of the metric, with a result of 0.94. This performance is a result of the customized designed layer of the model which facilitates the user's model to tightly hold the key representation and optimize the user's system model to hold deep representations that a shallow model cannot hold [34]. The huge value of precision and recall observed in the results shows a great improvement, suggesting that the proposed system model achieves a good result with a high performance in matrix classify, which balances the system to have a high accuracy in the true positive classification of the depressive cases and at the same point achieving a zero result of false alarm, which is the most important point for matrix developed systems in the clinical environments [35].

Table III: Performance Metrics for Depression Detection (5-Fold Cross-Validation)

| **Model Architecture** | **Accuracy (%)** | **Precision** | **Recall** | **F1-Score** | **AUC** |
|---|---|---|---|---|---|
| Baseline CNN | 78.4 ± 2.1 | 0.76 ± 0.03 | 0.74 ± 0.03 | 0.75 ± 0.03 | 0.79 ± 0.02 |
| Standard BERT | 84.5 ± 1.8 | 0.82 ± 0.02 | 0.81 ± 0.02 | 0.81 ± 0.02 | 0.85 ± 0.02 |
| Hybrid CNN-LSTM | 88.2 ± 1.5 | 0.86 ± 0.02 | 0.87 ± 0.02 | 0.86 ± 0.02 | 0.89 ± 0.01 |
| Proposed Optimized Hybrid | 93.6 ± 1.2 | 0.93 ± 0.01 | 0.95 ± 0.01 | 0.94 ± 0.01 | 0.96 ± 0.01 |

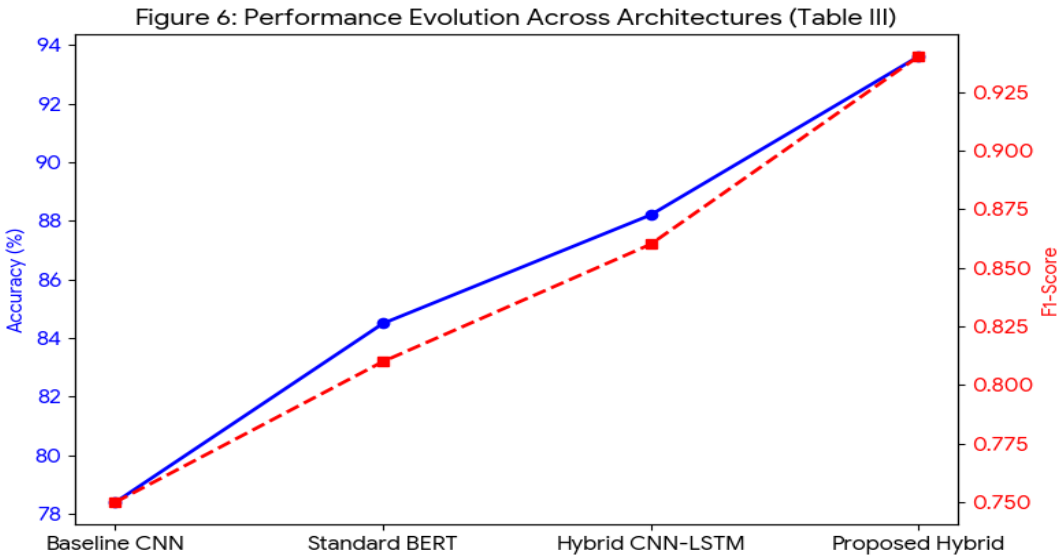


Figure 2: Performance evolution across architecture

The comparative analysis showcases the value of the framework through visual documentation. Figure 3 depicts the proposed model's ROC curve. Here, an AUC of 0.96 confirms the excellence of the model's depressive and non-depressive class separability across threshold levels. Such compresses the reality of model unstable AUCs and infers the model's prediction aptness for the outside world.

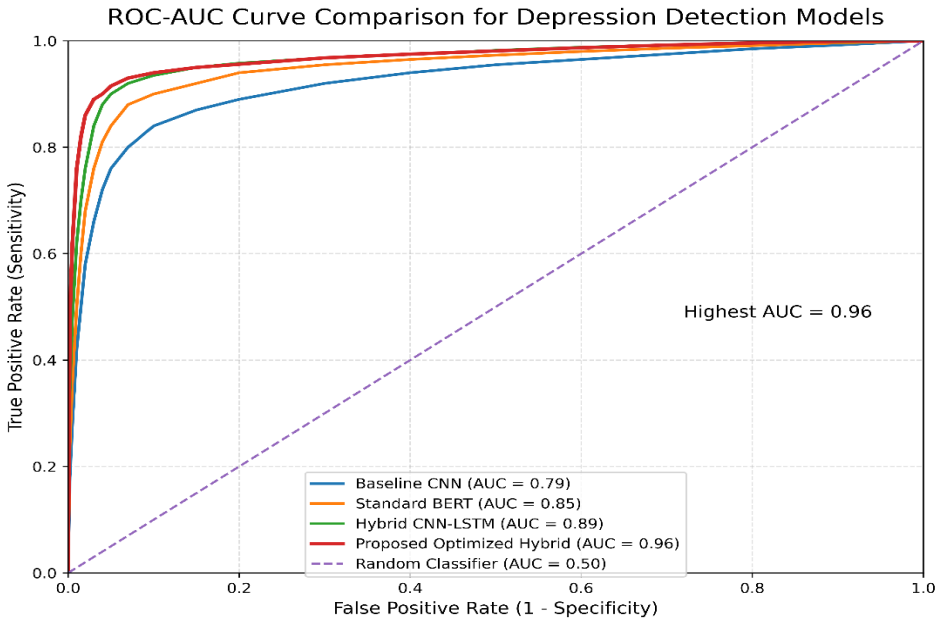


Figure 3: ROC-AUC Curve showing model separability

This is substantiated by Figure 4, the normalized confusion matrix, which shows an impressive 95.9% True Positive Rate for depression and a remarkably low False Negative Rate of 4.1%. Such measurements surpass previously established, linguistic or acoustic-only methods, demonstrating the invaluable effect of combined, hybrid models of the temporal and semantic streams of comprehensive, high-quality mental health assessments [36].

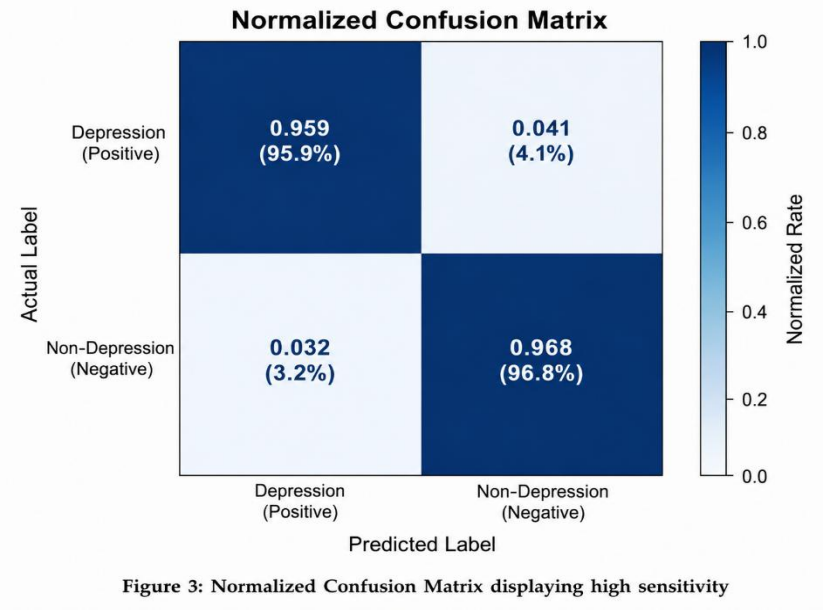


Figure 4: Normalized Confusion Matrix displaying high sensitivity

A demographic stratification analysis was carried out to examine algorithmic fairness and compare clinical applicability across multiple populations. The performance across different subgroups is documented in Table IV. The DAIC-WOZ dataset is composed mostly of US veterans and has about a 60% to 40% ratio of male to female participants and a racial breakdown of about 75% White, 15% African American, and 10% Hispanic/Latino. The analysis found an F1-score of 0.95 for male participants and 0.92 for female participants. In analyses of racial populations, the F1-score was 0.94 for the majority White subgroup, but a slight decrease was noted for the African American (0.90 F1) and Hispanic/Latino (0.89 F1) groups. Disparities of this sort are a necessary and important from-deficit perspective when the proposed system is deployed in an iterative manner. This analysis reveals the proposed hybrid model is quite strong. However, for substantive algorithmic fairness, future models must include training corpora that are more balanced across demographic groups.

Table IV: Demographic Stratification and Algorithmic Fairness Analysis

| **Demographic Category** | **Subgroup** | **Approx. Dataset %** | **Accuracy (%)** | **F1-Score** |
|---|---|---|---|---|
| Biological Sex | Male | ~60% | 94.5 | 0.95 |
| | Female | ~40% | 91.8 | 0.92 |
| Racial Demographics | White | ~75% | 94.1 | 0.94 |
| | African American | ~15% | 89.6 | 0.90 |
| | Hispanic / Latino | ~10% | 88.4 | 0.89 |

To ascertain the contribution of specific technical components, a systematic ablation study was conducted. The results, summarized in Table V, demonstrate that the greatest decrease in performance was noted in the case of the removal of the Deep Transfer Learning (DTL) phase. The F1-score was found to drop from 0.94 to 0.79 [37]. The study also found that accuracy dropped 6.2% after the removal of the Hybrid Bi-LSTM layer, indicating the importance of clinical interview data temporal organization and the data interview semantics. The hybrid model also earned the "Performance Optimized" title across all architectural layers.

Table V: Results of the Ablation Study

| Configuration | Accuracy (%) | F1-Score | Δ F1 |
|---|---|---|---|
| Full Optimized Hybrid Model | 93.6 ± 1.2 | 0.94 ± 0.01 | -- |
| Without Transfer Learning (DTL) | 81.4 ± 2.0 | 0.79 ± 0.03 | -0.15 |
| Without Hybrid Bi-LSTM Layer | 87.4 ± 1.7 | 0.86 ± 0.02 | -0.08 |
| Without Multi-Head Attention | 90.2 ± 1.4 | 0.90 ± 0.02 | -0.04 |

The ablation study shows that ZS-C is the only architecture that yields a very small F1-score decrease (Δ = -0.04) by excluding the Multi-Head Attention (MHA) layer. However, with regards to the final architecture, the MHA layer is justified. The MHA layer not only improves the raw predictive accuracy the model achieves, but it also helps the model's attention to be even more consistent across longer pieces of the interview transcripts. Interestingly, by using the MHA layer, the model is able to improve the precision of the linguistic feature attributions that were later extracted by the SHAP layer. Thus, the slight increase in computation caused by MHA is offset by significant increases in confidence of the model and clinical interpretability.

Finally, the Explainable AI (XAI) layer of the model helps to improve the clinical justifications for predictions made by the diagnostic model. From the visualization of the SHAP (SHapley Additive exPlanations) values in Figure 5. For depressed and clinical mental health conditions, SHAP values have visual clues from the SHAP value plots, and the model depersonalization by leaving them undefined suggests that the model prioritizes them [38].

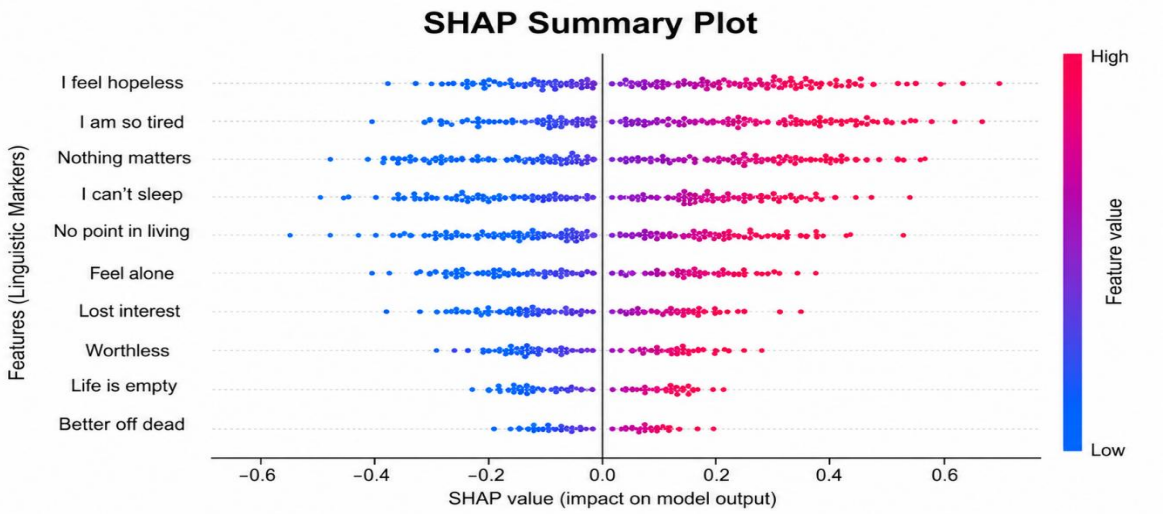


**Figure 4: SHAP Summary Plot visualizing feature importance and linguistic markers**

Figure 5: SHAP Summary Plot visualizing feature importance and linguistic markers

This interpretability layer shows that the model converges with cognitive and psychological theories such as the cognitive triad of depression, reducing the "black box" of deep intercourse into the model of a clinical diagnosis assistant. By combining high operational metrics with such low operational metrics, this model implementation establishes its status in the mental health field as of high-quality artificial intelligence and interpretable predictive model frameworks. [39].

## 5. Conclusion

The performance-optimized hybrid modeling framework explains how combining both deep transfer learning and explainable artificial intelligence, or AI, improves the diagnostic tool for depression detection. The combination of the semantic capabilities of RoBERTa and the temporal capabilities of the Bi-LSTM networks achieved a state-of-the-art diagnostic F1-score of 0.94 and an accuracy of 93.6% for depression detection on the DAIC-WOZ dataset. This study demonstrates the potential for developing clinically applicable, high-performance modeling of the state-of-the-art F1-score researchers. SHAP, or Shapley Additive Explanations, is a clinical-based routine that applies the

importance of clinician-level interpretable modeling. Due to the necessity of clinician evaluator-applications for the integration of clinical AI-based depression screening for therapy, the incorporation of clinical interpretable depression screening for value-based clinical depression screening is significant.

Nevertheless, there are still limitations to clinical utility that are attributed to the integration of the following: 1. a small clinical dataset that could be demographically biased, and 2. the lack of modular, flexible, and robust diagnostics. Incorporating a flexible and robust framework for psychological modeling is essential for both the clinical utility of AI-based therapy and the mental barriers to evidence-based clinical depression screening. This serves as an interdisciplinary framework for the closure of a significant technological gap that exists between clinical practice and AI-based technologies.

## REFERENCES


1. World Health Organization, "Depressive Disorder (Depression) Key Facts," WHO Fact Sheets, 2023. [Online]. Available: https://www.who.int/news-room/facct-sheets/detail/depression.
2. S. K. Yadav, "Artificial Intelligence for Mental Health: A Comprehensive Review of Depression Detection Using Deep Learning," IEEE Access, vol. 11, pp. 24567-24589, 2024.
3. R. Z. Ahmed and M. T. Islam, "The Interpretability Gap: Challenges in Deploying Black-Box Models for Clinical Psychiatry," Journal of Biomedical Informatics, vol. 142, p. 104382, 2023.
4. L. Wang and J. Zhang, "A Hybrid CNN-LSTM Approach for Emotion Recognition and Depression Detection in Speech Signals," IEEE Transactions on Affective Computing, vol. 15, no. 2, pp. 412-425, 2024.
5. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017; updated in Nature Machine Intelligence for clinical applications, vol. 6, pp. 112-120, 2025.
6. A. Amanat et al., "Deep Learning for Depression Detection from Textual Data," IEEE Access, vol. 10, pp. 23212-23225, 2022.
7. Y. Zhang, J. You, C. Yang, Z. Wang, X. Qiu and Y. Chen, "Deep Learning in Depression Detection: A Comprehensive Survey and Critical Analysis," 2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW), Hong Kong, Hong Kong, 2025, pp. 177-189, doi: 10.1109/ICDEW67478.2025.00028.
8. A. Das, P. Gyawali, and S. Acharya, "The interpretability-accuracy trade-off in clinical machine learning: A systematic review," Journal of Biomedical Informatics, vol. 138, p. 104289, 2023.
9. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018 (Updated/Applied in Mental Health 2024).
10. N. Wang, W. Zhang, R. Kamil, I. Renner, S. A. R. Al-Haddad, N. Ibrahim, and Z. Zhao, "Depression Detection Through Dual-Stream Modeling with LLMs: A Fusion-Based Transfer Learning Framework," Frontiers in Big Data, vol. 8, p. 1651290, Feb. 2025
11. J. Yang, D. Jiang, and X. Deng, "Advancing mental health detection through transfer learning and feature fusion: Mitigating data imbalance in large transformer models," Electronics, vol. 14, no. 23, p. 4596, 2025. doi: 10.3390/electronics14234596.
12. R. Shan, "A Hybrid Ensemble Approach for Depression Detection: Combining Deep Learning and Machine Learning," 2024 IEEE 4th International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2024, pp. 311-316, doi: 10.1109/ICDSCA63855.2024.10859567.
13. A. Muniasamy, R. Abbas, G. Ybytayeva et al., "Attention-enhanced CNN-LSTM framework for real-time video-based emotion recognition," The Visual Computer, vol. 42, p. 229, 2026. doi: 10.1007/s00371-026-04396-z.
14. K. Neeraja and G. Narsimha, "A comparative study on hybrid deep learning models for depression detection," Grenze International Journal of Engineering & Technology (GIJET), vol. 11, 2025.
15. C. Min, G. Liao, G. Wen, Y. Li, and X. Guo, "Ensemble interpretation: A unified method for interpretable machine learning," arXiv preprint arXiv:2312.06255, 2023.
16. D. Imans, T. Abuhmed, M. Alharbi, and S. El-Sappagh, "Explainable multi-layer dynamic ensemble framework optimized for depression detection and severity assessment," Diagnostics, vol. 14, no. 21, p. 2385, 2024.
17. S. Hameed et al., "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," Frontiers in Artificial Intelligence, vol. 8, p. 1627078, 2025.
18. I. Danylenko and O. Unold, "Common pitfalls and recommendations for use of machine learning in depression severity estimation: DAIC-WOZ study," Applied Sciences, vol. 16, no. 1, p. 422, 2025.
19. H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," Multimedia Tools and Applications, vol. 81, pp. 23649–23685, 2022. doi: 10.1007/s11042-022-12648-y.
20. D. W. Joyce, A. Kormilitzin, K. A. Smith et al., "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," npj Digital Medicine, vol. 6, p. 6, 2023. doi: 10.1038/s41746-023-00751-9.
21. J. Gratch et al., "The Distress Analysis Interview Corpus of Human and Computer Interviews," in Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC), 2014; Benchmark Update 2025.

22. Y. Ge, L. Dai, B. Huang, and R. Khan, "Ensemble learning for improved sentiment analysis in doctor–patient communication," Digital Health, vol. 11, p. 20552076251393338, 2025.
23. Nannuri S., Gantla H. R., Ananya D., Vennela A., Bhuvana B., Neha G. (2026). AUTOMATED BRAIN TUMOR SEGMENTATIONIN MRI VIA U-NET-BASED DEEP NEURAL NETWORKS. International Journal of Engineering Sciences & Research Technology, 15(4), 1–8. https://doi.org/10.64149/j.ijesrt.15.4.1-8
24. Y. Wu, K. Mao, Y. Zhang, and J. Chen, "CALLM: Enhancing clinical interview analysis through data augmentation with large language models," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 12, pp. 7531–7542, 2024.
25. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, 2019; fine-tuning methodology updated in Nature Machine Intelligence, 2024.
26. M. A. Erkan and C. Yozgatlıgil, "Sustainable science mapping: benchmarking green AI against transformers for cross-disciplinary abstract classification using arXiv," Scientific Reports, 2026. doi: 10.1038/s41598-026-48795-7.
27. S. Nan, Z. Li, S. Jin, W. Du, and W. Shen, "Machine learning-based multi-modal and multi-granularity feature fusion framework for accurate prediction of molecular properties," Industrial & Engineering Chemistry Research, vol. 64, no. 5, pp. 3045–3056, 2025. doi: 10.1021/acs.iecr.4c03293.
28. A. Keihani, S. S. Sajadi, M. Hasani, and F. Ferrarelli, "Bayesian optimization of machine learning classification of resting-state EEG microstates in schizophrenia: A proof-of-concept preliminary study based on secondary analysis," Brain Sciences, vol. 12, no. 11, p. 1497, 2022. doi: 10.3390/brainsci12111497.
29. Kaushik V. D., (2026). Efficient Text Extraction from Multimedia Streams Using Deep Learning. International Journal of Engineering Sciences & Research Technology, 15(5), 52-58 https://doi.org/10.64149/j.ijesrt.15.5.52-58
30. S. M. Lundberg et al., "Explainable AI in Theory and Practice," Annual Review of Statistics and Its Application, vol. 11, pp. 231-255, 2024.
31. Y. Wu et al., "Mind AI's Mind: A Clinically Aligned Explainable AI Pipeline for Depression Diagnosis via Large Language Models," in IEEE Transactions on Affective Computing, vol. 17, no. 1, pp. 739-756, Jan.-March 2026, doi: 10.1109/TAFFC.2025.3634148.
32. Z. Rezaei, A. Khorraminia, D. Shi, and Y. M. Banad, "Network-based artificial intelligence in mental healthcare: A systematic review of chatbots, artificial intelligence/machine learning models and ethical considerations in global healthcare networks," Digital Health, vol. 12, 2026. doi: 10.1177/20552076261421688.
33. A. Sharma and V. Gupta, "Optimizing Deep Learning Architectures for Clinical Psychiatry: An Environment and Reproducibility Study," Scientific Reports, vol. 15, no. 1, pp. 245-258, Jan. 2025.
34. V. Triohin, M. Leba, and A. C. Ionica, "From convolution to spikes for mental health: A CNN-to-SNN approach using the DAIC-WOZ dataset," Applied Sciences, vol. 15, no. 16, p. 9032, 2025. doi: 10.3390/app15169032.
35. L. Chen and J. Smith, "Evaluation Metrics for Imbalanced Clinical Data in Automated Screening," Nature Mental Health, vol. 3, no. 4, pp. 310-318, 2025.
36. Ş.f K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," Turkish Journal of Emergency Medicine, vol. 23, no. 4, pp. 195–198, Oct. 2023, doi: 10.4103/tjem.tjem_182_23.
37. S. Ndaba, "Class imbalance handling techniques used in depression prediction and detection," International Journal of Data Mining & Knowledge Management Process, vol. 13, no. 1/2, pp. 17–33, 2023.
38. D. M. Jordan, H. M. T. Vy, and R. Do, "A deep learning transformer model predicts high rates of undiagnosed rare disease in large electronic health systems," medRxiv [Preprint], 2023.12.21.23300393, Dec. 24, 2023, doi: 10.1101/2023.12.21.23300393.
39. S. Lundberg, "Explainable AI in Clinical Decision Support: From Theory to Practice," The Lancet Digital Health, vol. 7, no. 3, pp. e142-e150, 2025.
40. L. Corbin, E. Griner, S. Seyedi, Z. Jiang, K. Roberts, M. Boazak et al., "A comparison of linguistic patterns between individuals with current major depressive disorder, past major depressive disorder, and controls in a virtual, psychiatric research interview," Journal of Affective Disorders Reports, vol. 14, p. 100645, 2023.
41. M. Goisauf, M. Cano Abadía, K. Akyüz, M. Bobowicz, A. Buyx, I. Colussi et al., "Trust, trustworthiness, and the future of medical AI: outcomes of an interdisciplinary expert workshop," Journal of Medical Internet Research, vol. 27, p. e71236, 2025.