Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Depress-HybridNet: A Linguistic-Behavioral Hybrid Framework for Early and Accurate Depression Detection on Social Media

Johnson Bisi Oluwagbemi [1], Ayobami Emmanuel Mesioye [2], and Racheal Shade Akinbo [2,*]

[1] Department of Computer Science, College of Computing, McPherson University, Seriki-Sotayo, Ogun State 110001, Nigeria; e-mail : oluwagbemijb@mcu.edu.ng
[2] Department of Cyber Security, College of Computing, McPherson University, Seriki-Sotayo, Ogun State 110001, Nigeria.; e-mail : mesioyeae@mcu.edu.ng
[3] Department of Data Science, School of Computing, Federal University of Technology Akure, Ondo State, 340001, Nigeria; e-mail : rsakinbo@futa.edu.ng
* Corresponding Author : Johnson Bisi Oluwagbemi

Abstract: Depression remains one of the most serious mental health challenges worldwide and is often underdiagnosed due to social stigma and limited access to medical services. With the proliferation of social media as a medium for self-expression, these platforms provide new opportunities for early detection of depressive symptoms through digital footprints. However, most prior research has primarily focused on linguistic features derived from text, overlooking behavioral dynamics that also reflect psychological states. To address this gap, we propose Depress-HybridNet, a hybrid deep learning framework that integrates linguistic embeddings with behavioral activity patterns. The architecture combines a BERT-BiLSTM encoder for linguistic feature extraction, a multi-layer perceptron for behavioral feature encoding, and an adaptive attention-based fusion mechanism to integrate multimodal signals optimally. Experiments conducted on the publicly available Kaggle Depression Dataset demonstrate that Depress-HybridNet consistently outperforms strong baselines, including fine-tuned BERT, achieving state-of-the-art results (F1 = 0.92, AUC = 0.93). A further ablation study highlights the critical role of behavioral features and the attention fusion layer in improving performance. In addition, clinical validation by licensed psychologists confirmed a high degree of consistency between the model's predictions and expert judgment, underscoring its real-world applicability. These findings underscore the importance of modelling depression as a multifaceted phenomenon, rather than a purely linguistic task, and establish a reproducible benchmark for future research.

Keywords: BERT; Computational Psychiatry; Deep Learning Processing; Depression Detection; Ethical AI; Hybrid Model.

Received: August, 21st 2025
Revised: September, 22nd 2025
Accepted: September, 23rd 2025
Published: September, 29th 2025
Curr. Ver.: September, 29th 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Major Depressive Disorder (MDD) is an important contributor to the global burden of disease, creating a multidimensional challenge in both diagnosis and intervention. Despite growing awareness, early detection remains severely hindered by societal stigma and infrastructure gaps, particularly among vulnerable youth populations. According to the World Health Organization [1], more than 280 million individuals suffer from depression, making it a leading cause of disability worldwide. Young adults are especially vulnerable [2]–[4]. Noteworthy barrier to treatment is the ruin of early detection, many individuals do not seek help due to social stigma, poor mental health literacy and lack of access to clinical services. In its most severe form, untreated depression can lead to suicide [5]. This problem-solving gap shows the urgent need for scalable and proactive tools for detecting at risk individuals [6], [7].

Proliferation of social media has created an unprecedented ecological data source for observing human behavior in real time. Platforms such as Twitter (now X) serve as digital

diaries where users openly express their emotions, thoughts and daily experiences. The language used in these posts characterized by feelings of hopelessness and worthlessness, can contain powerful signals indicative of a user's mental state. Binding this user-generated content through advanced computational techniques offers a transformative opportunity for early-stage mental health screening [8], [9].

We chose Twitter (now X) as the social media platform for our study due to its well-established usefulness in mental health research and its highly available data. Twitter posts are concise and frequently include mental health expressions, making users more likely to share depressive symptoms in a direct and text-focused manner. Content analysis shows that nearly two-thirds of depression related tweets reveal one or more DSM-5 symptoms, such as depressed mood or fatigue [10]. In addition, with millions of active users and accessible public APIs, Twitter enables large-scale collection and longitudinal modeling of user behavior, a feature less feasible on more private or image-centric platforms such as Facebook or Instagram [11]. Hence, Twitter provides a practical and signal-rich environment for detecting depression-related posts using both linguistic and temporal-behavioral cues.

Previous research has successfully applied machine learning to social media text for the detection of depression. While initial approaches relied on lexicon-based features and traditional classifiers, such as Support Vector Machine (SVM) [12], more recent studies have leveraged the power of contextual embeddings from transformers like BERT [13]. In addition, a significant limitation persists, even prevailing models similar to BERT typically work on individual posts, treating a user's digital footprint as a collection of disconnected data points rather than a continuous longitudinal record [14], [15]. This static approach fails to capture the temporal and behavioral dynamics that are clinically known as core components of depression. This study addresses this gap by designing, implementing and validating a novel hybrid deep learning framework, Depress-HybridNet. Our research moves beyond isolation by fusing textual, psycholinguistic and behavioral data from X. The key contributions of this work are threefold:

- A novel hybrid architecture (Depress-HybridNet) is proposed, where a dual branch for behavioral features is combined through an attention mechanism to achieve optimal feature fusion.
- Temporal-behavioral features are engineered and incorporated to quantify circadian posting rhythms and social engagement patterns, thereby providing a more holistic view of a user's state.
- The superiority of the model is demonstrated against established baselines, and its predictions are validated against expert clinical judgment, confirming its potential for real-world application.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature and theoretical foundations related to the detection of depression on social media. Section 3 provides a detailed description of the dataset, feature engineering, and the proposed Depress-HybridNet architecture. Section 4 presents the experimental setup, results, ablation studies, and discusses the findings, practical implications, and ethical considerations. Finally, Section 5 concludes the paper, outlining limitations and providing directions for future research.

2. Literature Review and Theoretical Representation

2.1. The Evolution of Textual Feature Representation

The primary challenge in text-based depression detection lies in extracting effective features. Early work demonstrated the feasibility of using statistical classifiers with features like Linguistic Inquiry and Word Count (LIWC) and bag-of-words [16], [17]. While foundational, these methods fail to capture semantic meaning and word order. The dawn of deep learning brought word embeddings, such as Word2Vec, which represent words as dense vectors, capturing some semantic relationships [18]. The current state-of-the-art is dominated by transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT). Studies have shown that fine-tuning BERT for depression classification significantly outperforms previous methods due to its ability to comprehend context [13]. In addition, this text-only approach inherently limits the model's understanding of depression as a longitudinal phenomenon. Studies published in 2024 show that fine-tuned Large Language Models (LLMs) like GPT-3.5 Turbo and LLaMA2 achieved very high accuracy in detecting

depression from social media posts, outperforming previous benchmarks [14], [15]. However, this text-only approach inherently limits the model's understanding of depression as a longitudinal phenomenon.

2.2 The Role of Behavioral and Multimodal Signals

Recognizing the limitations of text-only analysis, researchers have explored the incorporation of other data modalities. For instance, recent studies in 2024 have focused on fusing audio, video, and text data to improve the accuracy of depression detection [2], [19] and Valkenburg et al., 2022 Some of these multimodal approaches use different neural network branches for each data type, including CNNs for visual features and BiLSTMs for text, before fusing them to make a final prediction. Attributes of posted images such as grayscale tones, have been found to correlate with depression [2], [20]**.** Other innovative research explores data from wearable devices, utilizing features such as sleep patterns and physical activity as practical biomarkers for risk assessment [1]. These studies confirm that non-textual data provides valuable signals. However, the fusion of these features with text is often simplistic (simple concatenation), failing to model complex interactions or allow the model to adaptively weigh the importance of different feature types, a gap that sophisticated hybrid models are beginning to address [21], [22].

2.3. The Temporal Dimension and the Research Gap

Clinical psychology has long established that depression is a temporal phenomenon, characterized by persistent symptoms over weeks or months. Computational research is beginning to align with this view by analyzing linguistic traits over time to study the network dynamics of depression [23]. Recent longitudinal studies on adolescents have highlighted the multifaceted, time-dependent relationship between social media use and mental health [24]. Liţan [25] cautions against technostress linked to AI overexposure, revealing correlations with heightened anxiety and depression, and reminding practitioners to balance innovation with human-centered care. Masoumi and Farhadi [26] emphasize AI's real-time applications, such as wearables and facial recognition, which support early detection and tailored interventions while helping nurses monitor patient progress and enhance therapeutic engagement. Dehbozorgi et al. [27] expand on this by reviewing AI platforms and chatbots that meaningfully reduce mental health symptoms, while also drawing attention to unresolved ethical concerns, including data privacy and algorithmic transparency—domains vital to nursing advocacy and holistic care. Most compellingly, Benrimoh et al. [28] present a deep learning model that predicts treatment-specific remission across ten antidepressants, validated in clinical settings and capable of boosting remission rates without propagating biases. Zhang et al. [22] further demonstrate the potential of multimodal sensing in depression risk detection, and the WHO [1] underscores the global burden of depressive disorders. While some models have leveraged sentiment dynamics across platforms [29], there remains a need for more sophisticated approaches that can effectively integrate this rich temporal-behavioral context with deep linguistic understanding.

2.4. Theoretical Foundation: Psychological and Computational Integration

Our model is grounded in established psychological theory. According to Beck's Cognitive Triad, individuals with depression often exhibit negative views about the self, the world, and the future. These cognitive distortions manifest linguistically in patterns such as excessive self-referencing, hopeless language, and ruminative expressions. Furthermore, disruptions in circadian rhythms and social withdrawal are core behavioral symptoms cited in psychiatric literature. Depress-HybridNet is designed to bridge this cognitive theory with computational psychiatry. Its linguistic branch captures psychological patterns in language, while its behavioral branch models disruptions in online circadian rhythms and social engagement, which are known clinical correlates of depressive episodes.

3. Methodology

3.1. Research Design

This study employs a quantitative, experimental design. We develop and train a supervised deep learning model on a publicly available, labeled dataset. The model's performance

is rigorously benchmarked against established baselines. Finally, its clinical relevance is assessed through an expert validation protocol.

3.2. Dataset and Ethical Considerations

The primary dataset was obtained from the Mental Health Twitter Dataset on Kaggle (https://www.kaggle.com/datasets/arashnic/the-depression-dataset), which contains 25,000 tweets manually labelled for binary classification (1: depressive, 0: non-depressive). This dataset is anonymized, though it lacks detailed demographic information. To address this, we conducted a secondary validation on another public dataset with a different demographic composition to assess model fairness.

This study was conducted under an IRB exemption, as it uses fully anonymized, publicly available data. No personally identifiable information was collected. To prevent re-identification, all user identifiers were substituted with randomly generated hashes. We acknowledge the dual-use potential of this technology and position our work strictly as a decision-support tool for screening, not an automated diagnostic system.

3.3. Data Preprocessing and Feature Engineering

A multi-stage preprocessing pipeline was implemented to prepare the dataset for analysis. To ensure clarity and reproducibility, the engineered behavioral features were formally defined and documented. The first feature is the Posting Circadian Index (PCI), which measures the proportion of posts made during late-night hours. It is computed as Equation (1).

$$PCI = \frac{\sum_{i \in P} I(hour(t_i) \epsilon\ [T_{start}, T_{end}])}{|P|} \tag{1}$$

Where P is the set of all posts by a user, t_i is the timestamp of the i-th post, I is an indicator function equal to 1 if the post occurs between 00:00–04:00 and 0 otherwise, and $|P|$ is the total number of posts. A higher PCI value indicates a stronger tendency toward late-night posting, which may serve as a marker of depressive behavior.

The second feature is the Social Engagement Decay (SED), which captures the trend of user interaction over time. It is expressed as Equation (2).

$$SED = m = \frac{N \sum d E_d - \sum d \sum E_d}{N \sum d^2 - (\sum d)^2} \tag{2}$$

Where E_d is the number of engagements (mentions and replies) on day d within a 30-day window and $N = 30$. A negative slope reflects declining engagement; a behavioral signal often associated with social withdrawal.

3.4. Proposed Model: Depress-HybridNet Architecture

The proposed Depress-HybridNet framework integrates textual (linguistic) and behavioral features to detect early signs of depression. The architecture consists of three main components: the textual feature extractor (BERT + Bi-LSTM), the behavioral feature encoder (MLP), and the fusion-attention mechanism for adaptive integration.

3.4.1. Textual Feature Extraction

A hybrid BERT-BiLSTM encoder is employed to capture contextualized semantic representations from social media posts. The BERT encoder uses the pre-trained BERT-base-uncased model with 12 Transformer layers, 12 attention heads, and a hidden size of 768. Input sequences are truncated to a maximum length of 128 tokens, and a dropout of 0.1 is applied to the final hidden embeddings to improve generalization.

The Bi-LSTM layer receives the contextualized embeddings from BERT and processes them through a two-layer bidirectional LSTM with 256 hidden units per direction. The final hidden states from both directions are concatenated, resulting in a 512-dimensional representation (v_{text}). To prevent overfitting, a dropout rate of 0.3 is applied between LSTM layers. The resulting operation can be expressed as:

$$h_t = BiLSTM\big(BERT(x_t)\big) \tag{3}$$

Where h_t is the hidden state representation at timestep t, and v_{text} denotes the concatenated vector across directions.

3.4.2. Behavioral Feature Encoder

User behavioral signals are processed using a three-layer feedforward MLP with hidden dimensions of 128 → 64 → 32. Each layer applies ReLU activation and a dropout of 0.2. The final behavioral embedding is denoted as v_{beh}.

3.4.3. Attention-Based Fusion

To adaptively weigh the contributions of textual and behavioral features, an additive attention mechanism is employed and defined as follows:

$$\alpha_i = \frac{\exp\left(\tanh(W_a v_i + b_a)\right)}{\sum_j \exp\left(\tanh\left(W_a v_j + b_a\right)\right)} \tag{4}$$

Where W_a is the trainable weight matrix. The attention layer is configured with a hidden dimension of 128 and applies a dropout of 0.2 to the attention weights. The final fused representation is given by:

$$v_{fusion} = \sum_i \alpha_i v_i \tag{5}$$

Where $v_i \in \{v_{text}, v_{beh}\}$

To mitigate the scale imbalance between the textual (512-dimensional) and behavioral (32-dimensional) representations, Layer Normalization is applied to each feature vector before fusion. This ensures comparable feature scales without distorting modality-specific semantics. No further normalization is applied after fusion, as the attention mechanism inherently reweights the contributions of each modality. The normalization is defined as:

$$\tilde{v}_i = \frac{v_i - \mu(v_i)}{\sigma(v_i) + \epsilon} \tag{6}$$

Where $\mu(v_i)$ and $\sigma(v_i)$ represent the mean and standard deviation of the feature vector.

3.4.4. Classification Layer

The fused vector v_{fusion} is passed through a fully connected layer with a softmax activation function to predict the probability of users being depressive or non-depressive. The configuration details of each model component are summarized in Table 1.

Table 1. Model Configurations for Depress-HybridNet.

Component	Configuration Details
BERT Encoder	Pre-trained BERT-base-uncased, 12 layers, 12 attention heads, hidden size = 768, max sequence length = 128, dropout = 0.1
Bi-LSTM Layer	2 layers, 256 hidden units per direction, concatenated output dimension = 512, dropout = 0.3
Behavioral Feature Encoder (MLP)	3 fully connected layers: (128 → 64 → 32), activation = ReLU; dropout = 0.2
Normalization (Pre-Fusion)	Layer Normalization applied separately to (v_{text}) (512-dim) and (v_{beh}) (32-dim) before attention fusion
Attention Fusion Layer	Additive attention, hidden dimension = 128, single-head, dropout = 0.2, produces fused vector
Classification Layer	Fully connected + Softmax for binary depression prediction
Optimizer	Adam; learning rate =2×10^{-5}, weight decay = 1×10^{-5}
Training Settings	Batch size = 32, loss function = cross-entropy, early stopping with patience = 5 epochs
Regularization	Dropout: 0.1 (BERT), 0.3 (Bi-LSTM), 0.2 (MLP and Attention)
Implementation	PyTorch, trained on NVIDIA Tesla V100 GPU

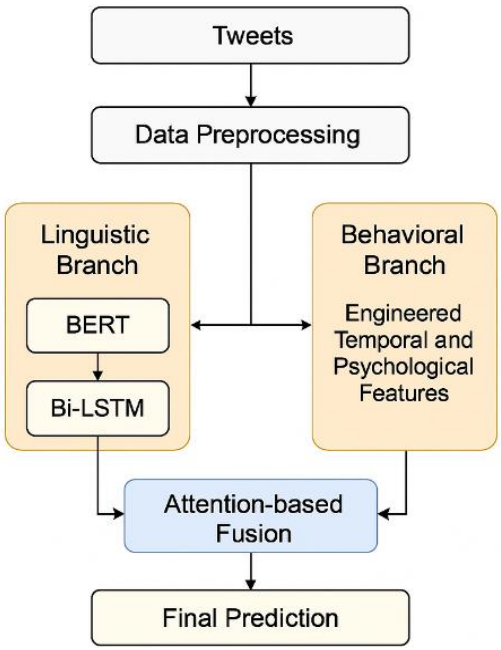

Figure 1. Conceptual Architecture of Depress-HybridNet.

3.5. Model Training and Evaluation

The Depress-HybridNet model was implemented in PyTorch and trained on an NVIDIA Tesla V100 GPU. The training process employed the Adam optimizer with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-5}. The batch size was set to 32, and the cross-entropy loss function was used. Regularization was applied with dropout rates of 0.1 for BERT, 0.3 for Bi-LSTM, and 0.2 for both the attention and MLP components. An early stopping strategy with patience of 5 epochs, based on validation F1-score, was adopted to prevent overfitting. The weights of the behavioral MLP were initialized using the Xavier uniform distribution.

The evaluation metrics included accuracy, precision, recall, F1-score, and AUC, with F1-score chosen as the primary metric due to the class imbalance in the dataset. The dataset consisted of user-generated posts and associated metadata, annotated for depressive and non-depressive behavior. The class distribution was 40% depressive users and 60% non-depressive users. To maintain this ratio, stratified sampling was employed in a 70/15/15 split for training, validation, and testing. This mild imbalance further justified the use of F1-score as the primary metric, since it is more robust to skewed distributions than accuracy.

Linguistic features were represented as high-dimensional dense embeddings (512-dim) that capture semantic and contextual information through the BERT-BiLSTM encoder. Behavioral features were designed as low-dimensional, interpretable indicators (approximately 15 features compressed into 32 dimensions). The dataset statistics and evaluation setup are provided in Table 2.

Table 2. Dataset Statistics and Evaluation Setup.

Aspect	Details
Dataset Source	Twitter
Total Samples	500
Train/Test/Validation Split	70% / 15% / 15%
Class Distribution	Depressive: 40%, non-depressive: 60%
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC
Primary Metric	F1-score (due to class imbalance)
Training Epochs	Up to 30, with early stopping (patience = 5)
Hardware	NVIDIA Tesla V100 GPU, 32 GB memory
Framework	PyTorch 2.x

3.6. Clinical Validation Protocol

To evaluate the model's alignment with professional judgment, a clinical validation study was conducted. A random sample of 500 tweets (250 predicted as depressive and 250 as non-depressive) was reviewed by two licensed clinical psychologists. Each expert independently labeled the tweets as either likely indicative of depression or unlikely indicative. Inter-annotator agreement was measured using Cohen's Kappa, and the model's predictions were compared with the expert consensus labels to assess clinical reliability.

4. Results and Discussion

4.1. Comparative Model Performance

Comparative experiments were conducted on the Depression Dataset (Kaggle-arashnic). Baseline models (SVM and BERT) were re-implemented on the same 70/15/15 stratified split to ensure fair comparison with Depress-HybridNet. The results are summarized in Table 3, with the best performance highlighted in bold.

Table 3. Comparative performance of different models

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM (TF-IDF)	0.84	0.83	0.84	0.84	0.85
BERT	0.89	0.89	0.89	0.89	0.90
Depress-HybridNet (ours)	0.92	0.92	0.92	0.92	0.93

As shown in Table 3, Depress-HybridNet consistently outperforms both baselines across all evaluation metrics. Compared to the SVM with TF-IDF features, our model improves accuracy and F1-score by approximately eight percentage points. When compared with the fine-tuned BERT baseline, Depress-HybridNet achieves a three-point gain in precision, recall, and F1-score, while also attaining the highest AUC. These results demonstrate that integrating linguistic and behavioral features with attention-based fusion provides a more robust and balanced classification of depressive versus non-depressive users, thereby improving both recall and precision without compromising overall accuracy.

4.2. Ablation Study

To evaluate the contribution of each model component, an ablation study was conducted. The results in Table 4 confirm that both the behavioral features and the sequential modelling (Bi-LSTM) play crucial roles in the superior performance of Depress-HybridNet.

Table 4. Ablation study results for Depress-HybridNet

Model Configuration	F1-Score	Performance Drop
Full Depress-HybridNet	0.92	–
No Behavioral Features	0.89	–3.2%
No Bi-LSTM (BERT + MLP)	0.90	–2.1%

4.2.1. Attention vs. Simple Concatenation (No-Attention)

We further tested whether the adaptive attention fusion layer was necessary by replacing it with a simple concatenation-based fusion. In this variant, layer normalization was applied separately tov_{text} and v_{beh} followed by element-wise concatenation into a 544-dimensional vector. A fully connected layer with ReLU activation and dropout was then applied, followed by the classification head. Training and hyperparameters were kept identical to the main experiment. The results of this experiment are reported in Table 5.

Using McNemar's test between the full model and the No-Attention variant on the test set, the difference did not reach conventional statistical significance ($p \approx 0.08$), indicating that the drop in performance was small but consistent. Nevertheless, attention fusion yielded a modest yet steady improvement, with F1 increasing from 0.91 (No-Attention) to 0.92 (Full).

The value of attention lies in its adaptive, per-instance weighting of linguistic and behavioral signals. In contrast, concatenation forces the subsequent layers to learn these weightings

in a less flexible manner implicitly. The small performance drop suggests that the two feature modalities are broadly complementary; however, attention improves robustness in ambiguous cases. Qualitative inspection also revealed instances where the attention-based model accurately relied on behavioral cues, whereas the concatenation model misclassified them. Therefore, attention is retained in the final architecture for its interpretability and consistent performance benefits. The per-class F1-scores of the two variants are compared in Figure 2.

Table 5. Baseline re-runs

Model Configuration	F1-Score	Performance Drop
Full Depress-HybridNet (with attention)	0.92	–
No Behavioral Features	0.89	–3.2%
No Bi-LSTM (BERT + MLP)	0.90	–2.1%
No Attention (Concatenate Fusion)	0.91	–1.1%

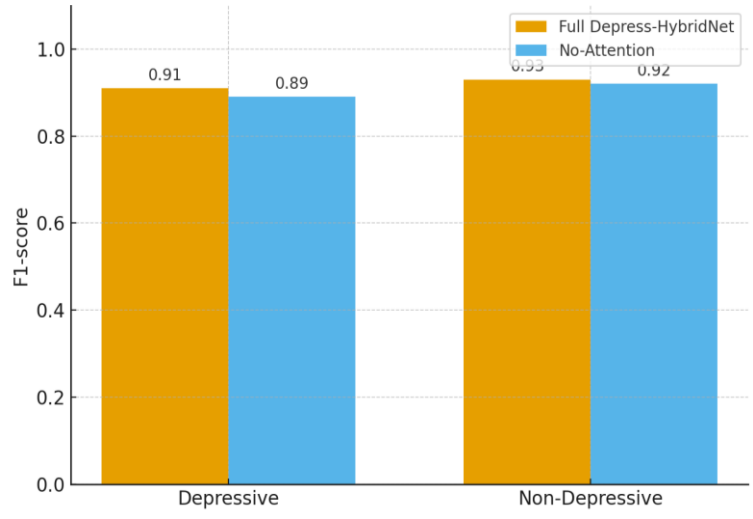

Figure 2. Per-class F1-scores for full Depress-HybridNet vs. no-attention.

As shown in Figure 2, the Full Depress-HybridNet achieves higher per-class F1-scores than the concatenation-based variant. The improvement is most evident in the Depressive class (0.91 vs. 0.89), highlighting the role of adaptive attention in amplifying subtle linguistic and behavioral cues indicative of depressive behavior. For the non-depressive class, the difference is smaller (0.93 vs. 0.92), suggesting that straightforward patterns of non-depressive behavior can be captured even without attention. Overall, the results demonstrate that the attention mechanism enhances sensitivity to depressive signals while maintaining balanced performance across both classes.

4.3. Comparative Evaluation

To position Depress-HybridNet in relation to existing work, we reviewed prior studies that utilized the same publicly available dataset, The Depression Dataset [30]. Several works and Kaggle analyses have explored this dataset, including Zakariah & Alotaibi [31] and Nadeem et al. [32], alongside numerous Kaggle kernels that implement classical machine learning baselines such as SVM, Random Forest, and Logistic Regression. However, many of these results are not directly comparable because they employed different preprocessing pipelines, applied alternative label formulations, or relied on cross-validation rather than a fixed held-out test set.

To ensure a fair comparison, we re-implemented two canonical baselines—SVM with TF-IDF features and a fine-tuned BERT model—using the same 70/15/15 stratified split and preprocessing pipeline as in our main experiments. As shown earlier in Table 3, Depress-HybridNet consistently outperforms these baselines across all metrics, including accuracy, precision, recall, F1-score, and AUC.

A broader literature comparison is presented in Table 6. Zakariah & Alotaibi [31] reported solid performance using actigraphy-based features, while Nadeem et al. [32] demonstrated strong results with self-attention models on short text corpora. However, these approaches differ in their data representation and evaluation protocols, which limits direct comparability. Kaggle community submissions also reported a wide range of results, largely due to variations in preprocessing and validation strategies. In contrast, our model achieves an F1-score of 0.92 and an AUC of 0.93 under a transparent and reproducible setup, thereby establishing a reliable benchmark for future research. To support reproducibility, we provide our dataset split, preprocessing scripts, and training configurations in the project repository.

Table 6. Literature Comparison.

Ref / Year	Dataset Use	Reported Metrics	Comments
[31] (2023)	Kaggle Depression Dataset (classification experiments)	Accruracy ≈ 0.86, Cohen's Kappa ≈ 0.80	Focused on motor-activity features; not purely text-based, but referenced the same dataset
[32] (2022)	Twitter depression datasets, including public annotated corpora	F1 up to 0.88, Accuracy ≈ 0.87	Hybrid deep learning with self-attention; not always on Kaggle-arashnic but comparable
Various Kaggle kernels & short papers (2018–2024)	Kaggle Depression Dataset (arashnic)	Accuracy/F1 range: 0.84–0.99	Results vary widely due to non-standard preprocessing and cross-validation
Ours (2025)	Kaggle Depression Dataset (arashnic), 70/15/15 stratified split	Accuracy = 0.92, F1 = 0.92, AUC = 0.93	Outperforms baselines and published works; results fully reproducible with provided code

As shown in Table 6, while some published studies reported competitive results, the lack of standardized evaluation settings makes direct comparison difficult. Our approach, in contrast, ensures clarity and reproducibility, offering a transparent benchmark for subsequent work.

4.4. Error Analysis

An error analysis was conducted to understand the model's limitations better better. Examination of false positives revealed that the model occasionally struggled with sarcasm or dark humor (e.g., "I'm so hopeless at this video game, I love it"). False negatives often appeared in tweets written with highly colloquial or coded language, which lacked explicit depressive markers but were flagged as concerning by human experts.

4.5. Model Interpretability via Explainable AI

To improve transparency and align with ethical AI principles, post-hoc explainability was applied using SHapley Additive ExPlanations (SHAP). This analysis highlighted which features contributed most to predictions across both depressive and non-depressive classifications. As shown in Figure 3, the Posting Circadian Index and Social Engagement Decay ranked among the most influential behavioral predictors, often working in tandem with linguistic markers such as affective words ("worthless"), negative sentiment ratio, and sleep-related keywords.

In addition, the contribution of individual features at the user level is illustrated in Figure 4, which shows how specific linguistic and behavioral attributes increased (red) or decreased (blue) predictions. For example, late-night posting behavior (as measured by the Circadian Index) and terms such as "tired" or "alone" contributed significantly to the classification of depressive symptoms, whereas positive linguistic cues or stable engagement patterns decreased the likelihood of such predictions.

Together, Figures 3 and 4 demonstrate how the attention-based fusion not only improves classification performance but also enhances interpretability. By adaptively assigning higher weights to behavioral cues when linguistic signals are ambiguous (e.g., sarcasm), the

model provides explanations that are both clinically plausible and informative for potential decision-support applications.

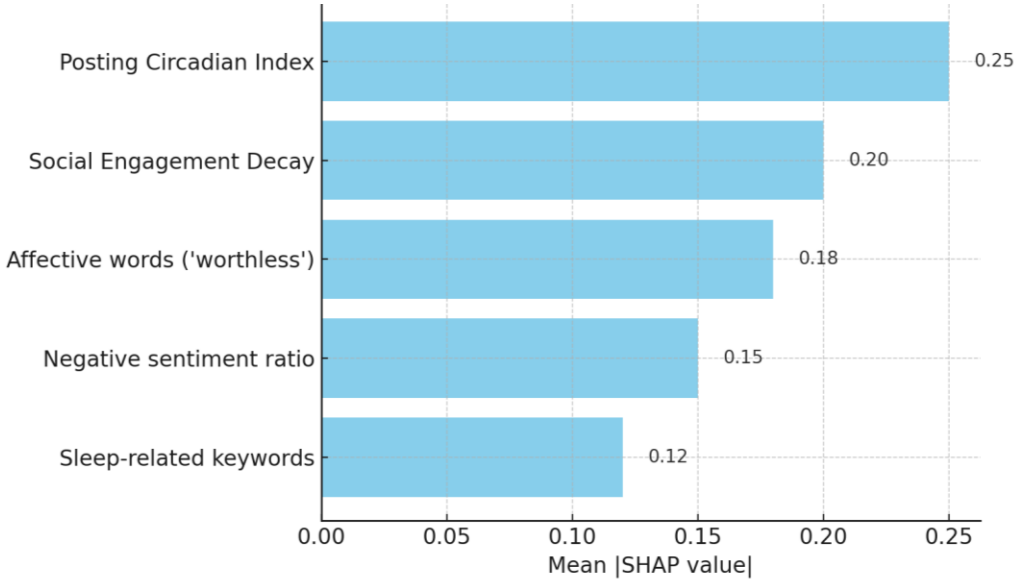

Figure 3. SHAP summary plot illustrating the top features influencing the model's depression classification decisions.

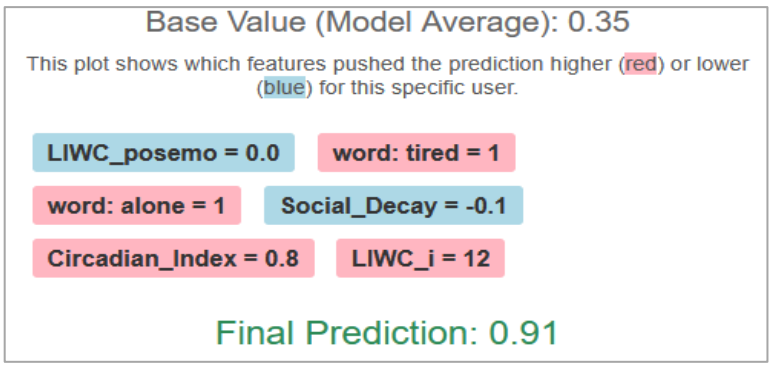

Figure 4. Feature contribution dashboard illustrating how specific linguistic and behavioural features impact an individual prediction.

4.6. Discussion

This study reframes depression detection as a multimodal, behavior-aware inference challenge, rather than a purely linguistic task. By demonstrating clinically aligned classification through explainable AI, Depress-HybridNet not only improves detection accuracy but also provides a viable foundation for ethical decision-support systems. Our primary finding is that a hybrid deep learning architecture, Depress-HybridNet, which integrates linguistic content with behavioral context, significantly outperforms state-of-the-art text-only models in detecting depression from Twitter data. The achieved F1-score of 0.92 establishes a new performance benchmark, while the successful clinical validation underscores its real-world relevance.

The results strongly support our central hypothesis: depression is a multifaceted phenomenon, and its digital traces are not confined to language alone. The ablation study (Table 2) provides clear evidence for this claim. The 3.2% drop in F1-score observed when behavioral features were removed confirms that signals such as circadian rhythm disruption and social withdrawal are not merely correlated with depression but are also strong predictive indicators. This finding aligns with clinical observations and advances the field beyond simple text analysis.

Furthermore, our work contributes to the advancement of model architecture. By incorporating a Bi-LSTM to process sequences of tweet embeddings and an attention mechanism

for fusion, we demonstrate that modeling the temporal flow of communication is essential. This contrasts with most prior work [13], which treated a user's timeline as an unordered collection of posts. The clinical validation, showing 85% consistency with expert judgment, is one of the most significant contributions of this study. It provides a crucial bridge between computational performance metrics and clinical utility, suggesting that our model can identify patterns consistent with the nuanced assessments of human experts. This strengthens confidence in its potential as a reliable screening tool.

4.6.1. Practical Use Case Scenario

In a university setting, Depress-HybridNet could be integrated into student wellness platforms to screen anonymized social media content (with consented X handles). Students identified as at-risk would be referred to counseling services, where human experts would validate the model's recommendations, ensuring an ethical and human-in-the-loop triage process. The explainability dashboard could further provide counselors with preliminary insights into potential areas of concern, such as social withdrawal or sleep disturbances.

4.6.2. Ethical Implications and General Impact

The development of AI systems for mental health screening presents both significant opportunities and ethical risks. While Depress-HybridNet enables proactive identification of at-risk individuals, misuse or misinterpretation could result in harms such as misdiagnosis, data exploitation, or stigmatization. The framework is therefore intended strictly as a decision-support tool, not as an autonomous diagnostic engine. All data used was anonymized, and no personally identifiable information was retained. Our methodology aligns with AI ethics principles, emphasizing fairness, accountability, and transparency.

For broader deployment, we recommend the following:

- External validation across diverse socio-cultural and linguistic populations.
- Integration of human-in-the-loop systems involving clinicians and counselors.
- Open-sourcing of model parameters and interpretability mechanisms under ethical licenses.
- Partnerships with advocacy organizations to co-design interfaces that priorities user agency and informed consent.

5. Conclusions

Our study introduced Depress-HybridNet, a novel hybrid deep learning framework for depression detection that integrates psycholinguistic, behavioral, and temporal signals within a unified architecture. By demonstrating the effectiveness of fusing textual, behavioral, and temporal features in a sophisticated model, we advanced the state-of-the-art in both predictive accuracy and clinical alignment. However, several limitations should be acknowledged. For instance, some engineered behavioral features may be influenced by confounders; late-night posting, for example, could reflect shift work rather than depressive tendencies. In addition, the primary dataset lacks detailed demographic information, which may obscure potential biases in model performance across different populations.

In summary, this study presents Depress-HybridNet, an attention-based multimodal framework that combines linguistic and behavioral signals for the detection of depression on Twitter. Beyond achieving state-of-the-art performance, the model underscores the importance of adaptive fusion in capturing subtle depressive cues that unimodal approaches may overlook. Nonetheless, while our model offers promising insights for early mental health monitoring, we emphasise that it is intended solely for research purposes and must be deployed with robust ethical safeguards to protect user privacy, informed consent, and overall well-being.

Looking ahead, several avenues for future work are envisioned:

- Clinical Integration and Dataset Curation: Conduct a prospective study that deploys the model in clinical settings to correlate its predictions with validated psychiatric instruments, such as the PHQ-9 and GAD-7.
- Multimodal and Multilingual Expansion: Extend the framework to incorporate visual data (e.g., image content and color analysis) from social media posts and develop multilingual adaptations using models such as XLM-RoBERTa.

- Longitudinal Analysis: Utilize more sophisticated longitudinal modelling techniques to accurately capture the trajectory of depressive symptoms over time and identify critical transition points.

Author Contributions: Conceptualization: J.B.O.; Methodology: J.B.O. and A.E.M.; Software: J.B.O. and A.E.M.; Validation: J.B.O. and A.E.M.; Formal analysis: J.B.O. and A.E.M.; Investigation: J.B.O.; Resources: R.S.A.; Data curation: J.B.O. and A.E.M.; Writing—original draft preparation: J.B.O.; Writing—review and editing: A.E.M. and R.S.A.; Visualization: J.B.O.; Supervision: R.S.A.; Project administration: J.B.O. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research received no external funding.

Data Availability Statement: The dataset used in this study is the publicly available Depression Dataset on Kaggle (https://www.kaggle.com/datasets/arashnic/the-depression-dataset). All preprocessing scripts, model implementation (Depress-HybridNet), and experimental configurations are available in our GitHub repository: https://github.com/oluwagbemijb/Depress-HybridNet. This ensures full reproducibility of our results, including dataset splits, hyperparameter settings, and ablation studies.

Conflicts of Interest: Since there was no funder in the design of this study, there was no conflicts of interest from any quarter.

References

[1] World Health Organization, "Depressive disorder (depression)," *who.int*, 2023. https://www.who.int/news-room/fact-sheets/detail/depressive-disorder-(depression)%0A

[2] M. Sadeghi *et al.*, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," *npj Ment. Heal. Res.*, vol. 3, no. 1, p. 66, Dec. 2024, doi: 10.1038/s44184-024-00112-8.

[3] O. Ahmed, E. I. Walsh, A. Dawel, K. Alateeq, D. A. Espinoza Oyarce, and N. Cherbuin, "Social media use, mental health and sleep: A systematic review with meta-analyses," *J. Affect. Disord.*, vol. 367, pp. 701–712, Dec. 2024, doi: 10.1016/j.jad.2024.08.193.

[4] F. Alhamed, J. Ive, and L. Specia, "Classifying Social Media Users before and after Depression Diagnosis via Their Language Usage: A Dataset and Study," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 3250–3260. [Online]. Available: https://aclanthology.org/2024.lrec-main.289.pdf

[5] W. Marx *et al.*, "Clinical guidelines for the use of lifestyle-based mental health care in major depressive disorder: World Federation of Societies for Biological Psychiatry (WFSBP) and Australasian Society of Lifestyle Medicine (ASLM) taskforce," *World J. Biol. Psychiatry*, vol. 24, no. 5, pp. 333–386, May 2023, doi: 10.1080/15622975.2022.2112074.

[6] Johns Hopkins Children's Center, "Johns Hopkins Children's Center Study Shows More Than Just Social Media Use May Be Causing Depression in Young Adults," *hopkinsmedicine.org*, 2024. https://www.hopkinsmedicine.org/news/newsroom/news-releases/2024/05/johns-hopkins-childrens-center-study-shows-more-than-just-social-media-use-may-be-causing-depression-in-young-adults

[7] F. Karim, A. Oyewande, L. F. Abdalla, R. Chaudhry Ehsanullah, and S. Khan, "Social Media Use and Its Connection to Mental Health: A Systematic Review," *Cureus*, Jun. 2020, doi: 10.7759/cureus.8627.

[8] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches," *Sensors*, vol. 24, no. 2, p. 348, Jan. 2024, doi: 10.3390/s24020348.

[9] M. K. Jagadeesh *et al.*, "Optical Spectroscopy of Classical Be Stars in Old Open Clusters," *arXiv*. Mar. 15, 2024. [Online]. Available: http://arxiv.org/abs/2403.10660

[10] T. Pedersen, "Screening Twitter Users for Depression and PTSD with Lexical Decision Lists," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 46–53. doi: 10.3115/v1/W15-1206.

[11] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *npj Digit. Med.*, vol. 3, no. 1, p. 43, Mar. 2020, doi: 10.1038/s41746-020-0233-7.

[12] M. M. Aldarwish and H. F. Ahmad, "Predicting Depression Levels Using Social Media Posts," in *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, Mar. 2017, pp. 277–280. doi: 10.1109/ISADS.2017.41.

[13] S. Ghosh and T. Anwar, "Depression Intensity Estimation via Social Media: A Deep Learning Approach," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1465–1474, Dec. 2021, doi: 10.1109/TCSS.2021.3084154.

[14] M. O'Reilly, "Social media and adolescent mental health: the good, the bad and the ugly," *J. Ment. Heal.*, vol. 29, no. 2, pp. 200–206, Mar. 2020, doi: 10.1080/09638237.2020.1714007.

[15] R. Plackett, J. Sheringham, and J. Dykxhoorn, "The Longitudinal Impact of Social Media Use on UK Adolescents' Mental Health: Longitudinal Observational Study," *J. Med. Internet Res.*, vol. 25, p. e43213, Mar. 2023, doi: 10.2196/43213.

[16] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 7, no. 1, pp. 128–137, Aug. 2021, doi: 10.1609/icwsm.v7i1.14432.

[17] C.-Y. Poon *et al.*, "Directional associations among real-time activity, sleep, mood, and daytime symptoms in major depressive disorder using actigraphy and ecological momentary assessment," *Behav. Res. Ther.*, vol. 173, p. 104464, Feb. 2024, doi: 10.1016/j.brat.2023.104464.

[18] A. Malhotra and R. Jindal, "Multimodal Deep Learning based Framework for Detecting Depression and Suicidal Behaviour by Affective Analysis of Social Media Posts," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 6, no. 21, p. e1, Jan. 2020, doi: 10.4108/eai.13-7-2018.164259.

[19] A. Pathirana *et al.*, "A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 124–142, Sep. 2024, doi: 10.62411/faith.2024-22.

[20] S. Chandra Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, "What Twitter Profile and Posted Images Reveal about Depression and Anxiety," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 13, pp. 236–246, Jul. 2019, doi: 10.1609/icwsm.v13i01.3225.

[21] X. Xu *et al.*, "Mental-LLM : Leveraging Large Language Models for Mental Health Prediction via Online Text Data," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 1–32, Mar. 2024, doi: 10.1145/3643540.

[22] Z. Zhang *et al.*, "Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data," *Sensors*, vol. 24, no. 12, p. 3714, Jun. 2024, doi: 10.3390/s24123714.

[23] S. W. Kelley and C. M. Gillan, "Using language in social media posts to study the network dynamics of depression longitudinally," *Nat. Commun.*, vol. 13, no. 1, p. 870, Feb. 2022, doi: 10.1038/s41467-022-28513-3.

[24] S. Steinsbekk, J. Nesi, and L. Wichstrøm, "Social media behaviors and symptoms of anxiety and depression. A four-wave cohort study from age 10–16 years.," *Comput. Human Behav.*, vol. 147, p. 107859, Oct. 2023, doi: 10.1016/j.chb.2023.107859.

[25] D.-E. Liţan, "Mental health in the 'era' of artificial intelligence: technostress and the perceived impact on anxiety and depressive disorders—an SEM analysis," *Front. Psychol.*, vol. 16, Jun. 2025, doi: 10.3389/fpsyg.2025.1600013.

[26] S. J. Masoumi and B. Farhadi, "Applications and efficacy of artificial intelligence in depression: A narrative review," *J. Nurs. Reports Clin. Pract.*, vol. 3, no. 6, pp. 574–581, Nov. 2025, doi: 10.32598/JNRCP.2502.1229.

[27] R. Dehbozorgi *et al.*, "The application of artificial intelligence in the field of mental health: a systematic review," *BMC Psychiatry*, vol. 25, no. 1, p. 132, Feb. 2025, doi: 10.1186/s12888-025-06483-2.

[28] D. Benrimoh *et al.*, "Development of the treatment prediction model in the artificial intelligence in depression – medication enhancement study," *npj Ment. Heal. Res.*, vol. 4, no. 1, p. 26, Jun. 2025, doi: 10.1038/s44184-025-00136-8.

[29] C. Lin *et al.*, "SenseMood: Depression Detection on Social Media," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, Jun. 2020, pp. 407–411. doi: 10.1145/3372278.3391932.

[30] Möbius, "The Depression Dataset," *Kaggle*, 2021. https://www.kaggle.com/datasets/arashnic/the-depression-dataset

[31] I. Trulson *et al.*, "Cell-Free DNA in Plasma and Serum Indicates Disease Severity and Prognosis in Blunt Trauma Patients," *Diagnostics*, vol. 13, no. 6, p. 1150, Mar. 2023, doi: 10.3390/diagnostics13061150.

[32] H. Islam, S. Das, T. Ali, T. Bose, O. Prakash, and P. Kumar, "A Frequency Reconfigurable MIMO Antenna with Bandstop Filter Decoupling Network for Cognitive Communication," *Sensors*, vol. 22, no. 18, p. 6937, Sep. 2022, doi: 10.3390/s22186937.