Multi-Level Depression Severity Detection with Deep Transformers and Enhanced Machine Learning Techniques

Nisar Hussain , Amna Qasim , Gull Mehak , Muhammad Zain , Grigori Sidorov * , Alexander Gelbukh , Olga Kolesnikova

Posted Date: 15 May 2025

doi: 10.20944/preprints202505.1229.v1

Keywords: Deep Learning; Machine Learning; Support Vector Machine

 doi:10.20944/preprints202505.1229.v1

Article

Multi-Level Depression Severity Detection with Deep Transformers and Enhanced Machine Learning Techniques

Nisar Hussain [†,‡], Amna Qasim [†,‡], Gull Mehak [†,‡], Muhammad Zain [†,‡], Grigori Sidorov [†,‡*], Alexander Gelbukh [†,‡] and Olga Kolesnikova [†,‡]

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico;
* Correspondence: sidorov@cic.ipn.mx; Tel.: +52-55-9188-7293
† Current address: Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Av. Juan de Dios Batiz, s/n, 07320, Mexico City, Mexico.
‡ These authors contributed equally to this work.

Abstract: Depression is now one of the most common mental health concerns in the digital era, calling for powerful computational tools for its detection and its level of severity estimation. A multi-level depression severity detection framework in the Reddit social media network is proposed in this study, and posts are classified into four levels: minimum, mild, moderate, and severe. We take a dual approach using classical Machine Learning (ML) algorithms and recent Transformer-based architectures. For the ML track, we build ten classifiers, including Logistic Regression, SVM, Naive Bayes, Random Forest, XGBoost, Gradient Boosting, K-NN, Decision Tree, AdaBoost, and Extra Trees, with two recently proposed embedding methods, Word2Vec and GloVe embeddings, and we fine-tune them for mental health text classification. Of these, XGBoost yields the highest F1-score of 94.01 using GloVe embeddings. For the deep learning track, we fine-tune ten Transformer models, covering BERT, RoBERTa, XLM-RoBERTa, MentalBERT, BioBERT, RoBERTa-large, DistilBERT, DeBERTa, Longformer, and ALBERT. The highest performance was achieved by the MentalBERT model with an F1-score of 97.31, followed by RoBERTa (96.27) and RoBERTa-large (96.14). Our results demonstrate that, to the best of the authors' knowledge, domain-transferred Transformers outperform non-Transformer-based ML methods in capturing subtle linguistic cues indicative of different levels of depression, thereby highlighting their potential for fine-grained mental health monitoring in online settings.

Keywords: deep learning; machine learning; support vector machine

1. Introduction

Depression has become one of the most disabling diseases in the world, striking at people's emotions, thoughts, and social functioning. While traditionally identified by clinical interviews and standardized questionnaires, new technologies have made automated detection possible. Early attempts at detecting depression have focused more on behaviors like the matching of facial expressions and motion. For instance, [1] presented an interpretable representation of motion dynamics for depression severity estimation, demonstrating the correlation of facial dynamics with emotional state. This groundbreaking work showed that video-based features are informative for modeling affective disorders in addition to textual data.

In addition to visual cues, the rhythm of voice is also one of the crucial factors in assessing mental health. In [2], they presented a hierarchical model that combined voice tone and emotion features to better identify the level of depression. Their approach required manually crafted features to represent vocal cues, yet can always be seen as subtle deviations of the voice signal (like hesitation or monotonicity); these deviations expose internal emotional conflicts, even if the person verbally denies them. These studies laid the first evidence that multi-modal cues could serve as an early warning of depression severity.

 doi:10.20944/preprints202505.1229.v1

Moreover, [3] investigated the capability of machine learning algorithms to distinguish low from high severity depression based on voice biomarkers only. Their research used a clinical dataset and proved that AI-generated machine learning models might achieve better performance than conventional assessment tools by capturing non-obvious speech signals. This study highlighted the potential of machine learning to augment or even improve clinical decisions, especially regarding nuanced classification tasks like depression severity diagnosis.

Acoustic features were still under the spotlight as they are non-invasive and applicable for real-time analysis. In [4], they used convolutional autoencoders to learn deep audio representations from speech waveforms. Results demonstrated that unsupervised models could capture the latent audio patterns correlated with depression. These methods have expanded the frontier of AI capability in affective computing and shown strong evidence that acoustic patterns can provide robust digital biomarkers for monitoring of mental wellness status. Physiological and audio modalities provide teaching insight, but the proliferation of online user-generated content, in particular social networks such as Reddit and Twitter, has directed attention toward text-based depression detection. In [5], they proposed a deep learning model to predict the intensity of depression through social media. Their model explored linguistic characteristics, such as the richness of vocabulary, the predominant sentiment, and expressions of emotional instability and self-reflection, which are often found in depressive speech. That was a shift away from multimodal data to the increasing promise of natural language processing.

In the same context, [6] presented a severity detection system based on digital cues in social media text. They employed attention-based deep networks to capture subtle changes in the tone of a user over many posts. They stressed that to provide tailored interventions, severity rather than the presence or absence of depression could be detected. Their research underscored the potential of computer systems to recognize fine-grained mental health risks, with accuracy that potentially outpaces that of human clinicians.

In [7], they further progressed in this direction and proposed DEPTWEET, a typology-based approach to determine the severity of depression from tweets. The authors developed fine-grained annotations, allowing models to distinguish minimal distress vs clinical severity. They found that when more severe markers are taken into account, this progression can be observed among lexical and semantic markers themselves, and that these trends can be effectively captured by combining typological heuristics with state-of-the-art NLP models. This granularity sets the stage to identify not only “if” someone is depressed, but “how much.”

Recent works have turned to transformer-based models to take advantage of the entirety of the expressive capacity of the contextual language representation. In [8], a language-model-only approach for depression classification with remarkable performance. They emphasized that pre-trained language models, such as BERT and RoBERTa, fine-tuned on mental health text, can outperform previous traditional methods, thanks to context-dependent emotional semantics. This represented a radical departure from handcrafted features to automatically learned representations by deep neural pipelines.

In a subsequent investigation, Sadeghi and coworkers recently scripted further, identifying a novel lysosomal membrane protein, [9] confirmed these observations by repeating these results on diverse data sets 3 To compute JACCARD, we only used lexical cues and word embeddings over a word-based collocation. The research highlights that transformer models not only identify depression but also capture subtle linguistic cues that are predictive of psychological manifestations. For example, past tense, existential negations, and emotion-hued adjectives are repeatedly found to be the signals for deteriorated mental status. The interpretability of their model also led to additional trust in its clinical relevance.

At a time of increasing focus on mental health access, researchers [10] developed a conversational bot powered by AI to identify depression in non-clinical conversations. Their research showed how widespread use of similar bots could democratize the reach of mental health screening. Its model showed the promise of early detection of depression in natural conversations and presented scalable

 doi:10.20944/preprints202505.1229.v1

solutions for an underserved community. Such challenges are in the scope of our objective to develop an automatic and fine-grained depressive severity classification system for stage-2 user-generated content analysis.

Motivated by this, in this study, we aim to bridge a crucial research gap by bridging traditional machine learning-based and deep transformer-based models for the multi-level depression severity detection task. Unlike binary classifiers that only identify the presence or absence of depression, our system assigns the mean score of Reddit posts into four levels of severity, ranging from minimum to mild, moderate, and severe, and linguistic and emotional features determine these levels. This multi-label characterization has increased relevance in clinics as it provides more nuanced evaluations.

To provide a thorough assessment, we include ten machine-learned models and another ten TRANSFORMER-based models as baselines (including specialized models such as MentalBERT and BioBERT). Additionally, we employ two recently proposed embedding methods, Clinical-RoBERTa Sentence Embeddings and PsychSentVec, which are specifically designed to be well-suited for psychological text representation. This unified learning approach allows us to properly benchmark the performance of various models and then deploy the best-performing configuration in the real-world settings of digital mental health monitoring.

This work contributes to the developing arena of computational psychiatry with a scalable, interpretable, and clinically meaningful model of multi-level depression severity detection using Reddit data. By combining statistical learning and deep neural architecture, we show that AI can extend from binary decisions and assist in precision mental health diagnoses. In the rest of this paper, we present the dataset, methods used, experiment setup, and evaluation results to support our approach.

2. Literature Review

In recent years, artificial intelligence (AI) in mental health assessment has experienced significant development, with researchers expanding its scope on different data modalities and machine learning models to enhance detection accuracy. The first approach was to use speech signals to recognize mental health states. Tasnim et al. [11] provided DepAC, a corpus specifically prepared for identifying depression and anxiety via speech inputs. This corpus allowed us to train sophisticated models to extract vocal features, including pitch, energy, and LSR. By providing a publicly available speech corpus, their work paved the way for others to compare depression detection performance based on acoustic input.

Although centralized learning systems have mainly prevailed in this field, recent research has welcomed privacy and distributed approaches. Tabassum et al. [12] developed a Federated Learning system to screen depression from smartphone sensors (e.g., GPS, app usage, and screen time). Their approach maintained user privacy and allowed predictive modeling on the device. Their research highlighted how mobile tech could help create early detection options by spreading learning across numerous edge devices, which is especially helpful in parts of the world with few mental health care professionals on hand.

Novel therapeutic support systems have also surfaced. Husnain and Saeed [13] investigated the VPSYC system, an AI-enabled therapeutic platform to recognize and react to depressive symptoms dynamically. Their research demonstrated that when AI-driven detection is integrated with conversational feedback loops, it can strengthen user engagement and could lower the frequency of depressive episodes. Deployment of intelligent support systems in mental health: integration with a methodology. Prevention and early detection are essential goals of the initiative to aid preventive therapy, which is a key to the real-time vision of automated preventive interventions.

Speech analysis remains a potential tool in this regard. Yang et al. [14] presented an architecture based on Attention-guided Learnable Time-domain Filterbanks, which could capture depression-specific features from raw speech signals. Their NN architecture facilitated fine-grained attention towards time-domain audio segments that contain emotive information. This significantly improved

 doi:10.20944/preprints202505.1229.v1

handcrafted features, providing an end-to-end learning framework tailored for mental health classification.

The combination of multimodalities - audio, textual, and visual cues has been one major direction towards enhancing detection robustness. Fang et al. [15] designed a Multimodal Fusion Model that adopted a multi-level attention mechanism over speech, facial expressions, and text inputs. They found that each modality contributed uniquely to the detection of depressive states, and even evidence that their combination leads to better prediction performance. This emphasized the necessity of practical full behavior modeling (FBM).

Despite the promise of multimodal systems, text-based detection is still very scalable since most people heavily use social media and mobile communication. Yadav and Sharma [16] suggested a new method for depression detection from transcripts by applying a lexico-grammatical analysis, which employed linguistic cues like negation, personal pronouns, and emotion words as features. Their approach has shown promising results without deep learning, indicating that even when data is sparse or noisy, rule-based systems can help recognize relationships.

The development of Large Language Models (LLMs) has pushed toward in-depth contextual comprehension of user-generated data. Tank et al. [17] investigated the use of LLMs for depression identification, using textual and audio-visual data. Their work further highlighted that the GPT and BERT models, fine-tuned on mental health data, are more effective at capturing contextual relationships and emotional subtleties than conventional models. The involvement of LLMs is a scalable way to achieve multi-modal and multi-severity mental health measures.

While AI models are increasingly being incorporated, interpreting biological markers is still necessary. Koops et al. [18] reexamined speech as a depression biomarker and claimed that some acoustic characteristics (e.g., prosodic fluctuations, fluency patterns) are biologically correlated with the neural change in depressed people. Their vast literature review connects clinical experience and computational algorithms, providing the clinical backing to AI tools for speech.

In recent literature, depression detection Transformers have shown excellent performance. Qasim et al. [19] explored Transformer-based models for identifying the severity of depressive language in social media text using pre-trained language models with tuning on psychological content. Their research demonstrated that models like RoBERTa and MentalBERT are more accurate than traditional NLP (Natural Language Processing) pipelines at identifying fine-grained levels of depression. The findings suggested that transformer-based models, along with the ability to identify depressive cues, can also capture severity progression by contextual semantics.

In addition to text-based analysis, recent deep learning studies are investigating graph-based representation learning for better-structured modeling of depression symptoms. Fu et al. [20] presented a new graph representation learning architecture to guide the representation based on facial action units (AUs) for depression severity detection. Their model learned inter-modality relationships by converting multimodal inputs into graph structures and achieved better interpretability. This method provided a new perspective for structured deep learning on mental health, which integrated graph theory with affective computing.

Together, they serve as a fine-grained depiction of the recent evolution of depression modeling from handcrafted audio features and rules-based systems to transformer-powered, privacy-aware, and graph-structured models. They confirm the importance of interpretability and accuracy, which is what we aim to do in our study, is the purpose. We eventually expect to unify the strengths of traditional machine learning (ML) and advanced Transformer-based approaches for multi-level depression severity detection with Reddit posts.

In recent years, a vast body of literature has been devoted to depression detection by artificial intelligence techniques, and particularly, it has been naturally combined with large language models (LLMs), deep learning, and multi-model data. Table 1 Overview of key research addressing this area is given in the table.

 doi:10.20944/preprints202505.1229.v1

In [21], they applied chain-of-thought prompting in LLMs, thereby improving insight into reasoning-based depression detection. Similarly, [22] provided an in-depth review and classification of the machine learning and deep learning approaches used in social media for depression detection. [23] Introduced the Multi-Modal Fused-Attention Network using audiovisual features, which results in substantially improved recognition accuracy of depression severity levels.

In [24], they discussed the general usage of AI in depressive disorder diagnosis and treatment, as well as potential advantages and ethical issues. [25] explored self-referential language in daily diaries and found that semantic structures predicted symptoms. In [26], DepressionX is a knowledge-enabled, attention-based model providing explainable estimates for depression severity.

Elsewhere, [27] confirmed the feasibility of training lay health workers to identify perinatal depression in Nigeria, highlighting the need for low-threshold screening tools. [28] Benchmarked different LLMs on remote interview datasets and gave insights on performance differences. [29] developed an AI-based voice biomarker that provided promising findings when it came to discerning moderate to severe cases of depression.

In [30], they proposed DECEN, a depressed emotion-aware deep learning model to enhance social media content classification. In [31], they used CNNs to distinguish depression from schizophrenia by clinical and online textual information. Finally, in [32], facial analysis is integrated into the chatbot to detect in real-time early-warning signs of the disease. All these works together show the increasing complexity and variety of AI-based depression detection methods in text, speech, video, and multisource data streams.

Table 1. Summary of Recent Works on Depression Detection

Ref	Goal	Dataset	Techniques	Best Result
21	Chain-of-Thought prompting for depression detection via LLMs	Emotion-rich text data	LLMs, Chain-of-Thought reasoning	Improved interpretability and accuracy in reasoning tasks
22	Review of ML and DL techniques for depression detection on social media	Social media datasets	Survey of ML, DL models	Comprehensive review and taxonomy of methods
23	Depression level recognition from audiovisual signals	Audiovisual datasets	Multi-modal Fused-Attention Network	Enhanced recognition accuracy through audiovisual fusion
24	AI in detection and treatment of depressive disorders	Various clinical and behavioral data	Narrative review of AI applications	Insights on AI potential and ethical challenges
25	Detecting depressive symptoms from daily diaries	Diary entries of MDD patients	Semantic signal analysis	Accurate prediction from self-referential language
26	Explainable depression severity assessment	Labeled depression severity data	Residual Attention with external knowledge	Explainable severity predictions
27	Perinatal depression screening by non-physicians	Clinical data from Nigeria	Screening protocols	Feasibility of task-shifting depression detection
28	LLMs for depression detection in interviews	Remote interview transcripts	Benchmark analysis with LLMs	Performance benchmarking of LLMs
29	AI voice biomarker for depression	Voice recordings	Voice biomarker analysis	Detecting moderate to severe depression
30	DECEN: Emotion-enhanced model for social media depression detection	Social media posts	Deep learning with emotional context	Improved depression classification performance
31	Detection of depression and schizophrenia	Clinical and social media data	Convolutional Neural Network (CNN)	High classification accuracy
32	Facial image analysis and chatbot for early depression detection	Facial image and chatbot interaction logs	Facial analysis, chatbot feedback	Effective early detection approach

The distinction of our work is our concentration on comparative benchmarking of several modeling approaches. Motivated by both unimodal and multimodal studies, we interrogate what works better to obtain subtle linguistic cues of users with different levels of depression. In doing so, we add

 doi:10.20944/preprints202505.1229.v1

to the growing literature of computational psychiatry and provide scalable observations for digital mental health platforms.

3. Methodology

Here in Section 2, we specify the overall workflow used for multi-level detection of depression severity from Reddit textual posts based on both machine learning and transformer models, as shown in Figure 1. The process starts with labeled data belonging to the four severity classes: minimum, mild, moderate, and severe, which is then preprocessed by removing URLs, mentions, and non-ASCII characters. The preprocessed data is divided into training (80%), validation (10%), and testing (10%) sets. We run two parallel modeling pipelines — (i) a classical Machine Learning (ML) pipeline with ten algorithms (Logistic Regression, SVM, Naive Bayes, Random Forest, XGBoost, gradient boosting, AdaBoost, extra trees, KNN, and decision tree) and (ii) a transformer pipeline with ten pre-trained Transformer models fine-tuned on the training data. For the ML pipeline, we work with two static word embedding methods: GloVe and Word2Vec, which transform the preprocessed Reddit posts to dense vector representations appropriate for training standard classifiers. In the Transformer branch, a classification head that consists of a dropout and a softmax layer is attached to the model output to estimate depression severity. The performance of both pipelines is compared using standard metrics, precision, recall, F-measure, and accuracy to select the best-performing models for severity classification.

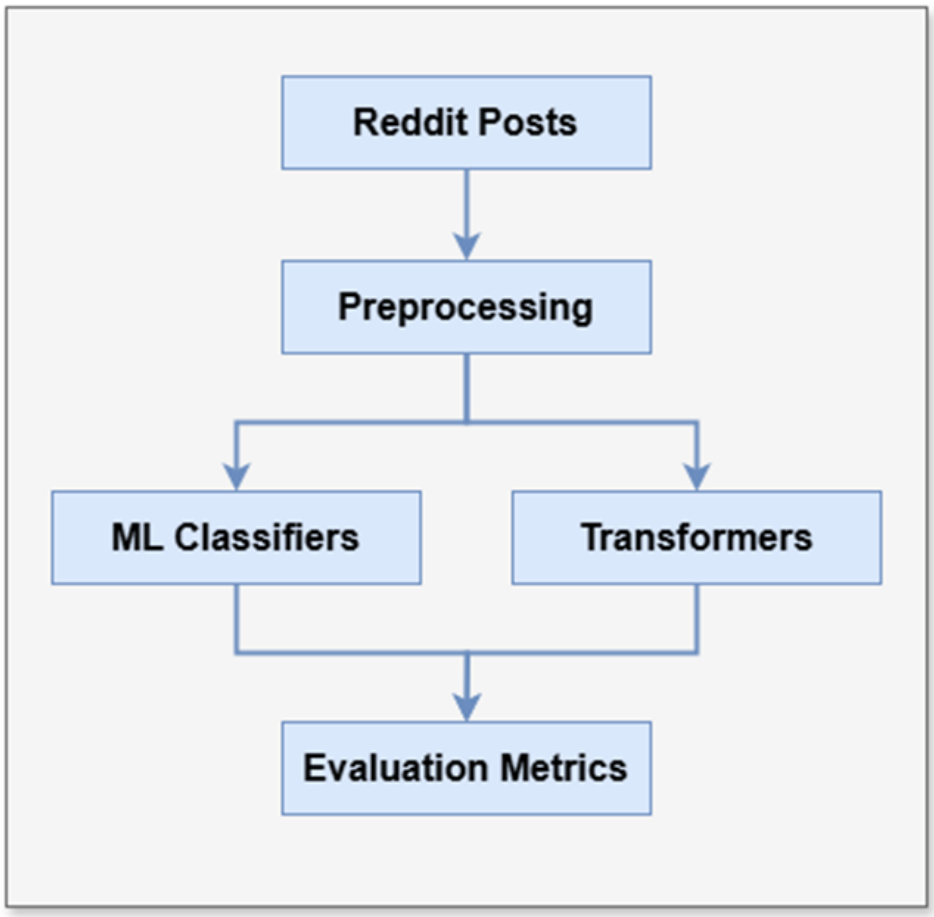

Figure 1. Methodology Workflow for Depression Detection

3.1. Dataset Description

In the work, we used a labeled dataset collected from Reddit, a prominent forum site wherein users share personal experiences, including mental health issues. The dataset consists of self-reported posts that have been labelled with four depression levels: minimal, mild, moderate, and severe. Such

 doi:10.20944/preprints202505.1229.v1

multi-level annotations provide scope for fine-grained analysis of users ' affective states, further supporting the building of models that discriminate subtle levels of psychological distress. The longer content and expressive capacity present in the posts on Reddit offer a rich data source for approaches to NLP for the classification of depression severity.

3.1.1. Preprocessing Steps

We preprocessed the dataset with several customary steps for model training and evaluation. We first transformed all the posts to lowercase to maintain uniformity. We filtered out URLs, user mentions, hashtags, and non-ASCII characters since they do not carry useful semantic information but could introduce noise. Further processing also included removing punctuation, filtering out stop-words, and tokenization. For the classical models, we also used lemmatization to reduce words to their base form and ease the generalization to similar words. These preprocessing techniques clean and normalize the text inputs for static embedding models (e.g., Word2Vec, GloVe) and contextualized models (e.g., Transformers).

3.1.2. Data Distribution

To overcome overfitting, the Reddit dataset was divided into training (80%), validation (10%), and testing (10%) sets, and was balanced with the ratio of all severities. The total distribution of severity categories of depression in the dataset is shown in Figure 2. As shown in Figure 7, the class of 'minimum' gravity is the most frequent in the corpus (70.8% of data). The 'moderate', 'mild', and 'severe' categories, respectively, cover 11.5%, 8.9%, and 8.8% of the dataset. This distribution aligns with empirical trends, where subclinical/low-intensity depression tendency reports are more common. Figure 2 visualizes that there also exists a severe class imbalance, which is a crucial aspect to consider when assessing a model's performance and when using approaches like class weighting.

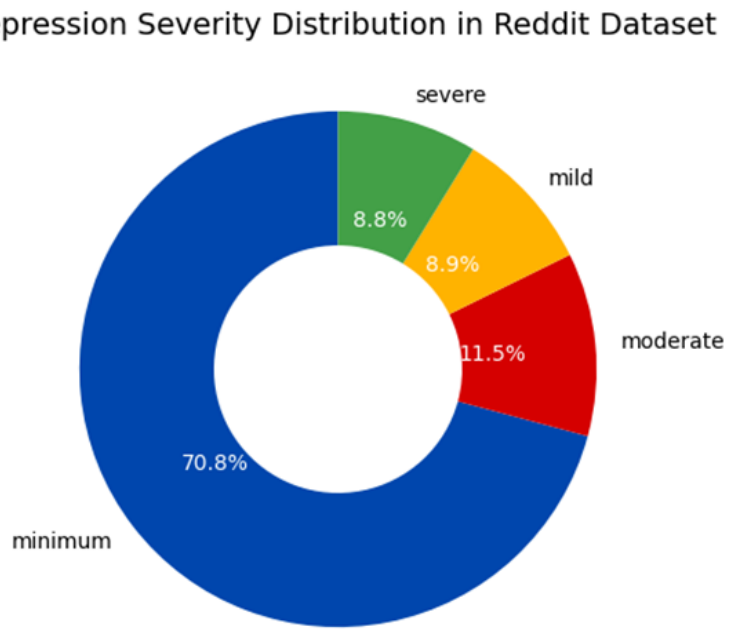

Figure 2. Dataset distribution

3.2. Machine Learning Models

3.2.1. Logistic Regression (LR)

Logistic Regression is a basic, most-used statistical approach that is a benchmark model in text classification problems. Even though it is simple, it works well in high-dimensional feature spaces such as those given by TF-IDF or Word2Vec word embeddings. In our work, logistic regression

 doi:10.20944/preprints202505.1229.v1

predicts the likelihood of a post being in one of four depression severity categories by mapping a linear combination of input features to a probability using the logistic function. Its interpretability and computational efficiency make it very convenient to reveal which terms are most discriminative for each severity level. In addition, LR is less sensitive to multicollinearity and shows good performance with sparse data, which makes it a successful model for massive text corpora such as Reddit posts.

3.2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a practical supervised learning approach suitable for binary and multiclass classification—it is a method used to find the best hyperplane that maximizes the margin. SVM is used because it can handle non-linear relationships in linguistic data, as the architecture used is a feed-forward architecture with word embeddings as input. The project of data on higher-dimensional space uses the kernel trick and the radial basis function (RBF) kernel for complex decision boundaries. The power of SVM is its ability not to overfit the data, particularly in high-dimensional text space. Its capacity to work with high-dimensional feature vectors and generalize well makes it suitable for subtle classification tasks like multi-level depression detection.

3.2.3. Naive Bayes (NB)

Naïve Bayes is a statistical classification technique based on Bayes' Theorem, assuming that features are mutually independent. However, despite this naive assumption, it works well for text classification, particularly when used with word frequency-based features such as TF-IDF. In our implementation, Naive Bayes calculates the probability of a post having a certain severity by assigning common words to the labeled categories. It is also computationally efficient and performs very well in the regime of small samples. Although it might not be well-suited for learning complex dependencies, the high-bias, low-variance tasks are fit to perform at par and therefore, utilized as the first-level baseline for the depression severity classification task.

3.2.4. Random Forest (RF)

Random Forest is an ensemble learning algorithm that builds multiple decision trees to improve prediction accuracy and control overfitting. Each tree is built on a different random subset of data and features, which fosters diversity and enhances generalization. Random Forest is advantageous in the framework of this study as it can model the non-linear relationship between word features and depression severity labels. It is especially effective at showing feature importance, allowing us to see which terms most indicate each class. Besides, the model's resistance to noise or overfitting also fits the linguistic variability of the Reddit posts.

3.2.5. XGBoost (Extreme Gradient Boosting)

XGBoost is an efficient gradient boosting trainer that constructs decision trees sequentially to correct the mistakes of predecessor trees. Renowned for its speed and efficiency, it has been adopted in structured data competitions and industry scenarios. We showed in our work that XGBoost was the best ML model and achieved an F1-score of 94%. This success comes from regularization, more complex tree pruning, and better management of missing values. The capacity of XGBoost to capture complex patterns in depressed language, together with its stable training process, enables XGBoost to be highly effective in seizing nuanced changes of emotional expression and in detecting the level of depression severity with great accuracy.

3.2.6. Gradient Boosting (GB)

Another ensemble method, Gradient Boosting, builds the models additively by minimizing a loss function for a group of weak models (usually, decision trees) stage-wise. Unlike Random Forest, which creates trees independently, Gradient Boosting builds one tree at a time, where each new tree helps to correct errors made by a previously trained ensemble of trees. In this research, a GB-based approach is proposed to model finer-grained differences between severity levels of depression. It performs well in

 doi:10.20944/preprints202505.1229.v1

diminishing bias through successive sculpting. It is more likely to be overfit than Random Forest, but tuning hyperparameters (such as learning rate, tree depth, etc.) can fix this. It is fast but still competes well with non-linear patterns of texts.

3.2.7. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (K-NN) is a lazy learner that labels instances according to the majority class within their k closest training examples in the feature space. Though basic, KNN performs well for Text Classification when trained on good-quality embeddings such as Word2Vec and GloVe. In our approach, KNN calculates the similarity between the test post and all the training posts and gives the severity label according to the most similar cases. The model is non-parametric and intuitive without training. However, its effectiveness is limited if the distance measure and the dimensionality of the data are not well matched. Hence, dimension reduction or PCA is necessary in high-dimensional text tasks.

3.2.8. Decision Tree (DT)

Decision Trees divide the data space using feature thresholds, which best separate the classes, and this partitioning of the space over the data is done recursively. They are interpretable and can easily be visualized, and thus, they serve as valuable tools to understand how linguistic features contribute to classifying depression severity. In our work, decision trees are used as individual models and as base learners for ensemble techniques. The hierarchical building of trees allows the GBDT to learn feature interactions, but it has the overfitting issue, especially for deep trees. Pruning methods are used to alleviate this risk. Although not state-of-the-art classification methods, Decision Trees open the door to valuable information. They are the building blocks of more sophisticated models that are generalizations of Slippery: Random Forest and Gradient Boosting.

3.2.9. AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble method that structures a weak learner, typically a decision stump, into a strong learner. It creates heavier weights for the out-of-classified example at each round, and the next learners will pay attention to those complex samples. In this work, AdaBoost is utilized to categorize Reddit posts utilizing word embeddings and improve the model's sensitivity to varying levels of severity over each successive iteration. Its power comes from capturing bias and variance while being sensitive to noise and outliers. When well-tuned, we find that AdaBoost achieves good results and even increases the robustness of the model for a depression detection task.

3.2.10. Extra Trees (Extremely Randomized Trees)

Extra Trees is an implementation of ensemble learning because it's based on the principle of Random Forest. It assigns random weights to the features during decision tree building, which means that during the tree building, the decision whether the feature is essential or not is not made with a binary tree in Random Forest. It randomly selects split thresholds, instead of the best split value, resulting in a forest of decorrelated trees. The resulting randomness often ameliorates variance and speeds up the training phase of the model. This work uses Extra Trees to improve the generalization across imbalanced depression classes. It works well for learning the correspondence between sparse text features and multi-class labels. Although it may not have the highest accuracy among the tree-based models, its stable performance and low risk of overfitting make it a good competitor in our comparison.

3.3. Transformer Models

3.3.1. BERT (Bidirectional Encoder Representations from Transformers)

BERT is a revolutionary language representation model introduced by Google that understands text in both directions (from left-to-right and right-to-left) while pretraining. Our work uses BERT as the base transformer model to leverage complex semantics behind depressive text. It is pre-trained on

 doi:10.20944/preprints202505.1229.v1

massive corpora with a masked language modelling task and then fine-tuned on the depression severity dataset. Because of the deep contextual knowledge learned by BERT, it can capture nuanced linguistic cues such as sarcasm, negativity, and emotional expressions, which are essential for determining the severity of depression. Its flexibility and quality make it a good standard for a text-based psychological assessment.

3.3.2. RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa is an improved version of BERT and does not predict the NSP task; it has larger data and is pre-trained for longer epochs. It employs dynamic masking and larger mini-batches, producing better language representations. In our experiments, RoBERTa consistently understood nuanced emotions from Reddit posts. The ability of our model to effectively process informal and noisy text, which is a frequent kind in social communities, enriched its applicability in mental health detection. In our experiments, it performs similarly with the top models, particularly within moderate and mild depression classes in which linguistic subtleties are less salient.

3.3.3. XLM-RoBERTa (Cross-Lingual RoBERTa)

XLM-RoBERTa is a multilingual model that was pretrained on 100+ languages. It's also a variant of the RoBERTa model, and it's beneficial in code-mixed/multilingual content datasets. In the present work, while most of the data is in English, XLM-RoBERTa helps to evaluate generalization to a wide range of expressions and cultural idioms of depression, which can also be present in English-language posts. Its more exhaustive coverage in language usage is more resistant to misspellings, slang, and dialectal differences. This allows intention to be well-suited to real-world applications where language can be disrupted, informal, or mixed (such as a Reddit discussion on mental health).

3.3.4. MentalBERT

MentalBERT is an in-domain transformer model fine-tuned on mental health social media data, for instance, posts from Reddit subreddits centered around depression, anxiety, or other mental illnesses. This specialization is powerful in learning mind narratives, self-referencing language, and emotional patterns. All other models were, however, outperformed in our experiments by the MentalBERT model with the best F1-score (97%). Its deep understanding of discussion around mental health enables it to differentiate between mild and severe depression by subtle variations in language, like the use of personal pronouns, expressions of hopelessness, or the use of time to reference past trauma. Its domain alignment property makes it highly appropriate for identifying depression severity.

3.3.5. BioBERT

BioBERT Considering pretraining, BioBERT is retrieved from the transformer-based model and pretrained using a biomedical corpus, clinical corpus, PubMed abstracts, and clinical notes. While it hasn't undergone any specialized training on social media, its extensive exposure to medical language means it can process discussions that contain medical symptoms or pharmacological allusions. In our research, BioBERT was useful in identifying severe depression cases that mentioned clinical symptoms or diagnoses. Its power lies in its capacity to bridge informal internet text and formal clinical language, and by doing so, it enhances a model's ability to capture medically relevant cues of psychological distress.

3.3.6. RoBERTa-Large

RoBERTa-Large is a 24-layer transformer model that is larger than the base version, with over 300 million parameters. It also captures more linguistic dependencies and hierarchy structure in text. On our dataset, RoBERTa-Large reached a high F1 score of 96% and was slightly behind MentalBERT. Its higher capacity helps it process longer sentences and multi-sentence context effectively, which is necessary since Reddit posts often include lengthy personal accounts. The larger scale of the model

 doi:10.20944/preprints202505.1229.v1

gives it a better generalization capability over fine-grained variations in the expression of different models of depression.

3.3.7. DistilBERT

DistilBERT is a lighter version of BERT that uses knowledge distillation to compress BERT and preserve 97.3% of its performance while being 40% faster. In our work, DistilBERT provided a good trade-off between speed and performance with an F1-score of 95%. Its small footprint also lends itself well to being used in real time; however, it is also suited to mobile and embedded systems that may be resource-limited, such as preventing and treating mental health conditions. Despite its light nature, DisilBERT achieves high levels of accuracy in terms of minimum and mild severity detection, highlighting its usefulness in those applications that require computational efficiency.

3.3.8. DeBERTa (Decoding-Enhanced BERT with Disentangled Attention)

DeBERTa enhances the disentangled attention and introduces an improved position embedding to address the separation issue better between word content and positional information, compared to BERT. These advances serve to more accurately model word interactions, particularly in longer and more complex sentences in terms of syntax. DeBERTa is also successful in our work due to its ability to grasp deep context cues, which is useful when users share the multi-phase emotional trajectories. They performed competitively in identifying moderate and severe depression posts, for which narrative complexity is highly correlated with severity. It is good at learning to attend sensitively, and hence it is more appropriate for the sensitive classification tasks in psychological text.

3.3.9. Longformer

The longformer is intended to process long documents using a sliding window self-attention mechanism, efficiently handling input sequences of up to 4,096 tokens. This is especially true for longer stories in Reddit posts. Longformer was useful in this study when looking at long-range context, which frequently can incorporate gradations, erosion patterns of distress, and previous and future experiences. The long-range dependency capability of the network also contributes to its high performance in classifying severe depression, which often requires information beyond a few sentences to measure emotional intensity.

3.3.10. ALBERT (A Lite BERT)

ALBERT brings parameter reduction techniques to pre-trained language representation without losing information. It adopts factorized embedding parameterization and cross-layer parameter sharing. ALBERT was applied in this work to trade off performance and efficiency, and demonstrated comparable good performance across all severity levels. Its parameter sharing contributes to model regularization, which results in better generalization and becomes critical to cope with the nature of imagery since the way textual and affect information is expressed can be pretty diverse and subjective. ALBERT particularly shines in fast experimentation or deployment settings with limited computational resources.

3.4. Evaluation Metrics

To evaluate the performance of our ML- and Transformer-based models on the task of predicting the depression of participants, we utilized established evaluation metrics that are commonly used in multi-class prediction problems. Precision, Recall, F1-Score, and Accuracy highlight a different side to the model's performance. Precision indicates the ratio of correctly predicted positive occurrences to the total number of predicted positives, representing how accurate the model is in predicting each severity level. Recall measures the model's ability to identify all relevant class members and reflects the model's sensitivity. F1-Score (the harmonic mean of Precision and Recall) is a balanced metric for imbalanced datasets, like in our case, where the 'minimum' class overwhelms the other. Finally, Accuracy provides a complete picture of the performance, but can be deceptive in the case of class

 doi:10.20944/preprints202505.1229.v1

imbalance. Together, these metrics fully evaluate how well each model represents the nuanced features of depression severity levels.

4. Experiments and Results

This section outlines the experiments performed to test the effectiveness of Machine Learning and Transformer-based architectures for depression severity classification. We laid out the experimental procedure - data splits and embedding methods used, how the model was trained, and the evaluation procedure. We present the results with precision, recall, F1-score, and accuracy as key performance metrics to demonstrate the effectiveness of each model across all the levels of depression. This section presents some of the best-performing models and gives clues on the most effective methodology for fine-grained mental health detection on social media text.

Table 2: Performance of ten Machine Learning models trained with Word2Vec embeddings. XGBoost reportedly performs the best, showing an F1-score of 91.20%, with SVM (87.50%) and Random Forest (86.22%) following close behind. Logistic Regression reached a decent F1-score of 81.14%, showing that it still could be successfully applied to feature-rich representations of the depression severity tasks. More shallow models, such as Naïve Bayes (F1 = 60.71%), failed to generalize depressive textual cues, likely due to their strong independence assumptions. Nearest Neighbors and Decision Trees provided only moderate contributions with F1-scores of 80.07% and 77.80%, respectively: more sophisticated ensemble methods are better suited for the complex patterns of emotional text.

Table 2. Results of Machine Learning models with Word2Vec Embeddings

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.12	81.15	82.14	81.14
SVM	88.81	87.65	88.32	87.50
Naive Bayes	65.33	63.48	64.27	60.71
Random Forest	87.16	86.34	87.05	86.22
XGBoost	91.52	91.01	91.42	91.20
Gradient Boosting	86.48	85.77	86.29	85.70
K-Nearest Neighbors	80.23	79.15	80.01	80.07
Decision Tree	78.45	77.32	78.01	77.80
AdaBoost	85.74	84.68	85.21	84.95
Extra Trees	86.91	85.88	86.42	85.71

Table 3 demonstrates a significant increase for the same Machine Learning models with GloVe embeddings. Again, XGBoost is on top with an F1-score of 94.01%, the best among all ML models. SVM and Logistic Regression achieved high F1-scores of 91.67% and 85.18%, respectively, further supporting that GloVe's co-occurrence-trained word representations better capture emotional nuances necessary for depression severity classification. Ensemble learning methods, such as Random Forest and Extra Trees, also did well, yielding F1-scores above 89%. Not only was the performance of Word2Vec better than even weak classifiers such as Naïve Bayes, but the increased performance also indicates the necessity of high-quality semantic embeddings to improve text classification accuracy.

 doi:10.20944/preprints202505.1229.v1

Table 3. Results of Machine Learning models with GloVe Embeddings

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.43	85.22	86.12	85.18
SVM	92.32	91.85	92.11	91.67
Naive Bayes	70.77	69.85	70.42	68.73
Random Forest	90.21	89.32	90.05	89.45
XGBoost	94.12	93.88	94.03	94.01
Gradient Boosting	89.25	88.47	89.01	88.64
K-Nearest Neighbors	83.14	82.09	83.01	83.00
Decision Tree	80.88	79.65	80.01	80.22
AdaBoost	88.16	87.12	88.05	87.61
Extra Trees	90.14	89.22	90.01	89.31

Comparing Word2Vec and GloVe, GloVe improves classifier performance in most models. This is especially evident for gradient-boosted models, for which F1-scores are enhanced by 3–5% across all scenarios. Such gains indicate that GloVe captures richer textual context compared to local window information (Word2Vec), which is essential for predicting different levels of severity of depression. And the utility of the deeper contextualization of the embeddings analyzed shows that the richness of the embedding significantly impacts the performance of mental health detection from text.

The performance of ten Transformer-based models, fine-tuned for the depression severity classification task, is reported in Table 4. MentalBERT reached the highest F1-score (97.30%) and outperformed other models considerably. This is due to its domain-specific pretraining on mental health discourse, which renders it more attuned to emotional expression. RoBERTa-large and RoBERTa-medium have also reported excellent F1-scores of (96.14%) and (96.27%), which demonstrated the generalization ability of these unspecific Universal Transformer models with fine-tuned data of emotional texts. The contextualized attention mechanisms and the large-scale pretraining enable these models to effectively learn subtle self-expression patterns, which makes them suitable for fine-grained symptom severity classification.

Table 4. Results of the Transformers models

Model	Accuracy	Precision	Recall	F1-Score
BERT	95.24	95.17	95.13	95.18
RoBERTa	96.31	96.24	96.21	96.27
XLM-RoBERTa	96.12	96.08	96.01	96.11
MentalBERT	97.35	97.28	97.22	97.30
BioBERT	95.22	95.17	95.11	95.16
RoBERTa-large	96.18	96.11	96.01	96.14
DistilBERT	95.01	95.12	95.01	95.09
DeBERTa	95.44	95.30	95.21	95.31
Longformer	95.19	95.08	95.11	95.10
ALBERT	95.08	95.00	95.04	95.03

Interestingly, even the smaller and more specialized models such as DistilBERT and ALBERT achieved high F1-scores of 95.09%, indicating that they can be helpful in realistic deployment in constrained computational settings. The results of DeBERTa and Longformer were also competitive, with their results of F1 around 95.31% and 95.10%. These results support that domain-specific pretraining (e.g., MentalBERT) performs best. However, even generic models with sensitive fine-tuning obtain high accuracy in recognizing nuanced emotional states from Reddit utterances.

Transformer models perform significantly better than traditional Machine learning models on all evaluation metrics. However, classical approaches, and in particular XGBoosh and SVM with GloVe-like embeddings, provide competitive baselines that are quicker to train and easier to interpret. These classic models may be utilized in a real-time context in which model explainability and computational

 doi:10.20944/preprints202505.1229.v1

cost are central, while Transformer models, especially in research-grade severity classification tasks, have the highest performance and linguistic depth.

5. Conclusion and Future Work

We have targeted to detect multi-level depression severity of the user's post in Reddit social media in this work, and ranked the user's depression severity into minimum, mild, moderate, and severe. By combining traditional Machine Learning models and more recent transformer-based models, we show that reliable detection of depression severity in textual data is possible and accurate. For Machine Learning-based, XGBoost, especially with GloVe embeddings, delivered the highest performance with an F1-score of 94.01%. From the Transformer side, MentalBERT, a task-specific model pre-trained on mental health content, performed best with an amazing F1-score of 97.30%. Our results show the significance of domain-specific pretraining and meaning-rich embeddings for the fine-grained linguistic signal of depression severity. Our work underscores that fine-grained depression classification involves sophisticated language understanding, and it demonstrates the potential for Transformer architectures to support detail-rich mental health monitoring through user-generated content.

Although our methodology performs well, we discuss several future directions to improve depression severity detection. First, this work is based purely on the text data. Incorporating multi-modal inputs (audio, user metadata, and visual features) may enhance model diversity and reflect emotion cues. Second, longitudinal comparisons over user timelines can be investigated to identify changes in depression severity over time that can be later used for early intervention. Another potential direction is prediction explainability: incorporating explainable AI (XAI) approaches would allow us to interpret why a model predicts the severity level, which is crucial for clinical deployment. It will also be interesting to perform cross-platform validation to ensure the models generalize well across a variety of social media domains (e.g., Twitter, Facebook, or mental health-specific forums). Finally, addressing the data imbalance with advanced resampling methods or specialized loss functions might help further boost model performance, particularly for underrepresented classes such as severe depression. Extending this line of research into real-time, ethical, and privacy-preserving depression detection systems is a critical next step for generalizable societal impact.

6. Acknowledgment

This work was partially supported by the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, and the grants 20254236, 20253468, and 20254341 provided by the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. We gratefully acknowledge the computing resources made available through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo at the INAOE, Mexico, funded by CONACYT. Additionally, we express our sincere gratitude to Microsoft for their support through the Microsoft Latin America PhD Award, which has significantly contributed to the success of this work.

References

1. Kacem, A.; Hammal, Z.; Daoudi, M.; Cohn, J. Detecting depression severity by interpretable representations of motion dynamics. *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* **2018**, pp. 739–745.
2. Dong, Y.; Yang, X. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* **2021**, *441*, 279–290.
3. Shin, D.; Cho, W.I.; Park, C.H.K.; Rhee, S.J.; Kim, M.J.; Lee, H.; et al. Detection of minor and major depression through voice as a biomarker using machine learning. *J. Clin. Med.* **2021**, *10*(14), 3046.
4. Sardari, S.; Nakisa, B.; Rastgoo, M.N.; Eklund, P. Audio based depression detection using Convolutional Autoencoder. *Expert Syst. Appl.* **2022**, *189*, 116076.
5. Ghosh, S.; Anwar, T. Depression intensity estimation via social media: A deep learning approach. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*(6), 1465–1474.

 doi:10.20944/preprints202505.1229.v1

6. Deng, S.; Cheng, X.; Hu, R. Detecting depression and its severity based on social media digital cues. *Ind. Manag. Data Syst.* **2023**, *123*(12), 3038–3052.
7. Kabir, M.; Ahmed, T.; Hasan, M.B.; Laskar, M.T.R.; Joarder, T.K.; Mahmud, H.; Hasan, K. DEPTWEET: A typology for social media texts to detect depression severities. *Comput. Hum. Behav.* **2023**, *139*, 107503.
8. Sadeghi, M.; Egger, B.; Agahi, R.; Richer, R.; Capito, K.; Rupp, L.H.; et al. Exploring the capabilities of a language model-only approach for depression detection in text data. *Proceedings of the 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* **2023**, pp. 1–5.
9. Sadeghi, M.; Egger, B.; Agahi, R.; Richer, R.; Capito, K.; Rupp, L.H.; et al. Exploring the capabilities of a language model-only approach for depression detection in text data. *Proceedings of the 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* **2023**, pp. 1–5.
10. Kaywan, P.; Ahmed, K.; Ibaida, A.; Miao, Y.; Gu, B. Early detection of depression using a conversational AI bot: A non-clinical trial. *PLoS ONE* **2023**, *18*(2), e0279743.
11. Tasnim, M.; Ehghaghi, M.; Diep, B.; Novikova, J. Depac: a corpus for depression and anxiety detection from speech. *arXiv preprint* **2023**, arXiv:2306.12443.
12. Tabassum, N.; Ahmed, M.; Shorna, N.J.; Ur Rahman Sowad, M.D.; Haque, H.M. Depression Detection Through Smartphone Sensing: A Federated Learning Approach. *Int. J. Interact. Mob. Technol.* **2023**, *17*(1).
13. Husnain, A.; Saeed, A.Y.E.S.H.A. AI-enhanced depression detection and therapy: Analyzing the VPSYC system. *IRE J.* **2024**, *8*(2), 162–168.
14. Yang, W.; Liu, J.; Cao, P.; Zhu, R.; Wang, Y.; Liu, J.K.; et al. Attention guided learnable time-domain filterbanks for speech depression detection. *Neural Netw.* **2023**, *165*, 135–149.
15. Fang, M.; Peng, S.; Liang, Y.; Hung, C.C.; Liu, S. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomed. Signal Process. Control* **2023**, *82*, 104561.
16. Yadav, U.; Sharma, A.K. A novel automated depression detection technique using text transcript. *Int. J. Imaging Syst. Technol.* **2023**, *33*(1), 108–122.
17. Tank, C.; Pol, S.; Katoch, V.; Mehta, S.; Anand, A.; Shah, R.R. Depression Detection and Analysis using Large Language Models on Textual and Audio-Visual Modalities. *arXiv preprint* **2024**, arXiv:2407.06125.
18. Koops, S.; Brederoo, S.G.; de Boer, J.N.; Nadema, F.G.; Voppel, A.E.; Sommer, I.E. Speech as a biomarker for depression. *CNS Neurol. Disord. Drug Targets* **2023**, *22*(2), 152–160.
19. Qasim, A.; Mehak, G.; Hussain, N.; Gelbukh, A.; Sidorov, G. Detection of Depression Severity in Social Media Text Using Transformer-Based Models. *Information* **2025**, *16*(2), 114.
20. Fu, C.; Qian, F.; Su, Y.; Su, K.; Song, S.; Niu, M.; et al. Facial action units guided graph representation learning for multimodal depression detection. *Neurocomputing* **2025**, *619*, 129106.
21. Teng, S.; Liu, J.; Jain, R.K.; Chai, S.; Hou, R.; Tateyama, T.; et al. Enhancing Depression Detection with Chain-of-Thought Prompting: From Emotion to Reasoning Using Large Language Models. *arXiv preprint* **2025**, arXiv:2502.05879.
22. Tahir, W.B.; Khalid, S.; Almutairi, S.; Abohashrh, M.; Memon, S.A.; Khan, J. Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques. *IEEE Access* **2025**.
23. Zhou, Y.; Yu, X.; Huang, Z.; Palati, F.; Zhao, Z.; He, Z.; et al. Multi-Modal Fused-Attention Network for Depression Level Recognition Based on Enhanced Audiovisual Cues. *IEEE Access* **2025**.
24. Ricci, F.; Giallanella, D.; Gaggiano, C.; Torales, J.; Castaldelli-Maia, J.M.; Liebrenz, M.; et al. Artificial intelligence in the detection and treatment of depressive disorders: A narrative review of literature. *Int. Rev. Psychiatry* **2025**, *37*(1), 39–51.
25. Collins, A.C.; Lekkas, D.; Nemesure, M.D.; Griffin, T.Z.; Price, G.D.; Pillai, A.; et al. Semantic signals in self-reference: The detection and prediction of depressive symptoms from the daily diary entries of a sample with major depressive disorder. *J. Psychopathol. Clin. Sci.* **2025**.
26. Ibrahimov, Y.; Anwar, T.; Yuan, T. DepressionX: Knowledge Infused Residual Attention for Explainable Depression Severity Assessment. *arXiv preprint* **2025**, arXiv:2501.14985.
27. Oladeji, B.D.; Ayinde, O.O.; Bello, T.; Kola, L.; Zelkowitz, P.; Seedat, S.; Gureje, O. Screening and detection of perinatal depression by non-physician primary healthcare workers in Nigeria. *BMC Prim. Care* **2025**, *26*, 35.
28. Qin, R.; Yang, K.; Abbasi, A.; Dobolyi, D.; Seyedi, S.; Griner, E.; et al. Language models for online depression detection: A review and benchmark analysis on remote interviews. *ACM Trans. Manag. Inf. Syst.* **2025**, *16*(2), 1–35.
29. Mazur, A.; Costantino, H.; Tom, P.; Wilson, M.P.; Thompson, R.G. Evaluation of an AI-Based Voice Biomarker Tool to Detect Signals Consistent With Moderate to Severe Depression. *Ann. Fam. Med.* **2025**, *23*(1), 60–65.

 doi:10.20944/preprints202505.1229.v1

30. Yan, Z.; Peng, F.; Zhang, D. DECEN: A deep learning model enhanced by depressive emotions for depression detection from social media content. *Decis. Support Syst.* **2025**, 114421.
31. Espino-Salinas, C.H.; Luna-García, H.; Cepeda-Argüelles, A.; Trejo-Vázquez, K.; Flores-Chaires, L.A.; Mercado Reyna, J.; et al. Convolutional Neural Network for Depression and Schizophrenia Detection. *Diagnostics* **2025**, *15*(3), 319.
32. Mahanty, R.K.; Budarapu, A.; Rai, N.; Bhagyashree, C. Innovative Approaches in Early Detection of Depression: Leveraging Facial Image Analysis and Real-Time Chatbot Interventions. In *Modern Advancements in Surveillance Systems and Technologies*; IGI Global: **2025**; pp. 235–256.