DeprexAI: An AI Based System for Early Detection and Support in Depression Analysis

Mohit Bohra
Department of Information Technology
ABES Institute of Technology
Ghaziabad,India
bohramohit93199@gmail.com

Nishant Tomer
Department of Information Technology
ABES Institute of Technology
Ghaziabad,India
nishanttomar008@gmail.com

Ritu Ranjan Chaubey
Department of Information Technology
ABES Institute of Technology
Ghaziabad,India
Chaubeyrituranjan5@gmail.com

Piyush Chauhan
Department of Information Technology
ABES Institute of Technology
Ghaziabad,India
peeyushrajput321@gmail.com

Paras Garg
Department of Information Technology
ABES Institute of Technology
Ghaziabad,India
parasg1609@gmail.com

Reshu Tyagi
Department of Information Technology
ABES Institute of Technology
Ghaziabad, India
reshu.tyagi0907@gmail.com

ABSTRACT— Depression is a most critical health issues nowadays in public health sector which is affecting more than 260 million people worldwide. Even though this issue is in attention still early detection remains difficult due to social shame and lack of awareness. Although there is also limited access to mental health professionals, but these social media platforms where users express themselves can serve as signs of mental health conditions.

We made DepreXAI an AI based model which helps to identify and analyse depression in three levels from tweets. We used three small transformer models ESG, ABV and XDL which gives results in three levels mild, moderate and severe individually and to increase the accuracy and robustness of the model we combine the outputs of all three models by a hybrid ensemble averaging method which gives a more accurate and precise results. We also used Explainable AI in which we used its two frame work which are SHAP and LIME. They help in justifying each results and increase the transparency of the model's output.

Keywords— Depression Detection, Transformer Models, Ensemble Learning, Explainable AI, Natural Language Processing, Social Media Analytics.

I. INTRODUCTION

Depression is a popular mental health condition worldwide which is affecting more than 260 million people around the world according to WHO. Due to social shame, it usually remains untreated because of lack of awareness and few mental health professionals. But in the growing digital era and digital footprints of users on digital platforms like twitter has helped to detect the early warning signs of depression through NLP techniques. In these two studies, CLPsych and eRisk they explained how these textual hints in social media posts can help in detecting depression signs [1], [2].

Recent development in NLP techniques and deep learning have enabled automated depression detection using pre trained transformer models like BERT, RoBERTa, and ALBERT [9], [13]. However, these models have beaten old machine learning methods but they act as black boxes for doctors and researchers as they provide less transparency for results [14], [15]. Also, these single model systems are likely to suffer on dataset specific bias and also generalization issues when used in diverse population [11], [12].

To solve these problems, we made DepreXAI a transformer-based ensemble model developed for explaining depression detection from social media text. We integrated three fine tuned lightweight transformers ESG, ABV, XDL whose outputs are averaged to give a more accurate final result. Also, we integrated the final model with Explainable AI using its two framework SHAP and LIME to increase reasonings and transparency for doctors and researchers [16], [17]. Our Project also supports real time monitoring of depression trends in region specific to give regional mental health reports and level to help policy makers to take steps accordingly [22], [24].

II. RELATED WORKS

Use of social media for depression detection is an active research for many years and its keep evolving from understanding text hints and statistical methods to transformer based models. Previous works like CLPsych (2015) and eRisk Lab(2017) gave benchmark datasets and frameworks for identifying and evaluating depression behaviors from tweets and reddit post [1], [2]. These work introduces the concept of early risk detection, where model analyze the posts to give results on mental state before the treatment [3], [4]. If we see other researches also explored the new ways to study the depression patterns from text posts to develop a base for interpretable depression classifiers.

Introduction of NLP in depression classification made a revolutionary success in this domain. Transformer models like BERT, RoBERTa, and ALBERT gave the superior performance in classifying depression signs because of their ability to understand contextual language [9], [11], [13]. Later on several studies between 2021 and 2024 extended these models to detect depression in levels like mild, moderate, severe using ICD -10 mood disorder classifications (F32/F33) [19], [20], [21]. However, even with high accuracy these models act as a black box system which made doctors and researchers unable to validate the results [14], [15].
Now to overcome this problem, they started using Explainable AI with mental health detection models. In which these frameworks LIME, SHAP and LRP have been used to study the text and its emotional features to affect the

model's decision [16], [17], [18]. Researches like Explainable AI for Mental Health (npj Digital Medicine 2023) and Toward XAI for Mental Health Detection from Language (Frontiers in Psychiatry 2023) showed how it help in building trust and transparency [14], [15].

Also, several researches also investigated how depression signs vary time to time which helped in visualizing emotional trends across different regions [22], [23], [24], [25]. These studies helped to find the potential for large scale real time health monitoring system which can help policy makers and organizing awareness campaigns.

While previous works have made a good progress but most depends on single model systems that may overfit to specific datasets. Whereas, DepreXAI is made of a hybrid ensemble of three small transformer models (ESG, ABV, XDL) combined through average based method, along with integrated XAI. This multi model explainable approach bridges the gap between predictive accuracy and transparency which provides a more reliable and interpretable solution for real-world mental health applications.

III. PROPOSED METHODLOGY

The proposed system, DepreXAI, is a deep learning–based ensemble framework designed for explainable depression detection using social media data. The methodology is divided into different stages: data collection from social media and annotation, preprocessing, model training, ensemble fusion of different models, and explainable AI integration for explaining the reason behind the output of the model. The complete architecture is explained in Fig. 1.

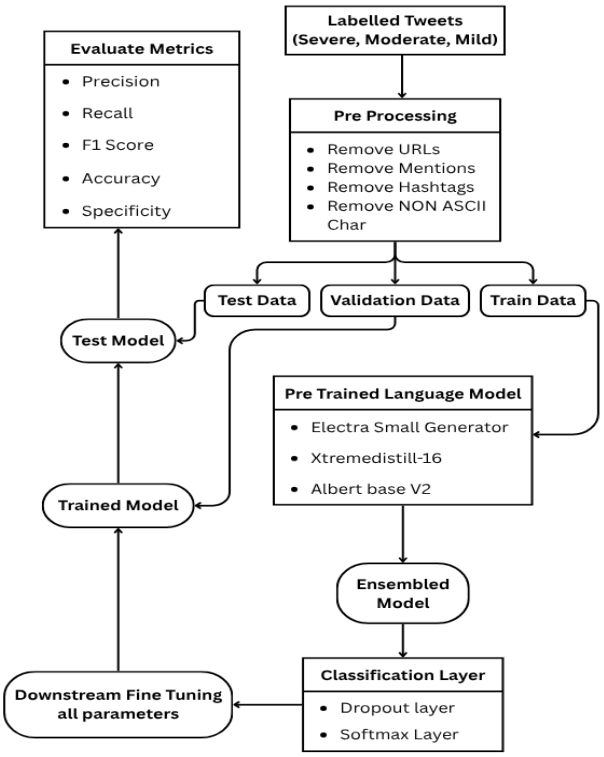

Fig. 1. Model architecture

A. Data Collection and Annotation

DepreXAI use publicly available Twitter datasets containing user regular posts and overview on different topics which are related to depressive expressions and emotional states. Tweets are taken from the Twitter API using domain-specific hashtags like *#depressed*, *#mentalhealth*, and *#feelingdown*. After the data acquisition, posts are filtered and anotate into three classes—mild, moderate, and severe depression—based on the ICD-10 diagnostic criteria (F32/F33) for mood disorders [19], [20]. Additional sentiment scores are derived using VADER and TextBlob to support semi-supervised annotation. The annotated dataset is then split into training (70%), validation (15%), and testing (15%) sets for model development.

B. Data Preprocessing

Preprocessing is used for removing the unwanted noise and standardize the input for the model development. Each tweet undergoes the following transformations:

- Removal of URLs, emojis, Hashtags, and user mentions used in the tweets.
- Conversion of words in lowercase letter and tokenization using the WordPiece tokenizer generally used for models like BERT, ALBERT, DistilBERT, Electra.
- Stop-word removal and lemmatization of words for getting the root word of the words used in the tweets.
- Padding and truncation to a maximum sequence length of 130 tokens for getting the same size of input.

This process helps us to get the consistent textual representation for transformer-based models and reduces bias introduced by informal online language in different social media platform like twitter(X), etc.

C. Transformer Model Training

Three transformer-based models—ESG, ABV, and XDL—are fine-tuned independently on the depression-labeled dataset taken from twitter(X). Each model is a smaller, more efficient version of a transformer neural network.

- ESG Model focuses on emotional sentiment analysis use domain-specific embeddings techniques to capture affective intensity of different words in the sentence.
- ABV Model add attention-based variance mechanisms to help in enhancing contextual feature extraction.
- XDL Model uses cross-domain linguistic pretraining to generalize across different demographic text patterns.

Each model gives outputs as softmax probability vector across the three depression levels. Model training is conducted using Adam optimizer with an initial learning rate of 2e-5, a batch size of 16, and early stopping is used to prevent overfitting. Evaluation metrics include accuracy, precision, recall, F1-score are used for evaluating the model effectiveness.

D. Ensemble Fusion Mechanism

To improve prediction stability of out models and minimize model bias, all the predictions of ESG, ABV and XDL are integrated via ensemble method. The ensemble joint output for the j-th depression class model is determined by:

$$P_{final} = \frac{P_{ESG} + P_{ABV} + P_{XDL}}{3}$$

This average outcomes of all three models taken which strengthen the robustness and generalize our models. The result is assigned a class label according to the maximum value in P_{final}.

E. Explainable AI (XAI) Integration

For the transparency and interpretability of predictions, DepreXAI applies Explainable AI (XAI) approaches such as SHAP (Shapley Additive Explaination) and LIME (Local Interpretable Model Explanation). These frameworks describe the influence of each input feature (word, phrase, or token) on the model's final decision, converting a complex deep network into an easy explanation space suitable for clinical understanding.

1) SHAP-Based Global Interpretability

SHAP is grounded in cooperative game theory, where each input feature is treated as a "player" contributing to the model's output.
For an instance x and model f, the SHAP value ϕ_i for feature i is computed as the average marginal contribution of that feature across all possible subsets S of features:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]$$

Where F denotes the full set of features. This quantifies how much the inclusion of feature i changes the model's output prediction. For textual data, each token embedding acts as a feature vector, and the SHAP library aggregates their contributions to display *positive* (red) and *negative* (blue) word influences on the depression severity class.

For each tweet x_j, the model output probability for the depressed class $f(x_j)$ can be expressed as:

$$f(x_j) = E[f(x)] + \sum_{\{i=1\}_{\{ij\}}^{\{n\}\setminus phi}$$

Where $E[f(x)]$ is the expected model prediction over the dataset and ϕ_{ij} represents the SHAP value contribution of token i in instance j. A higher positive indicates that the corresponding word (e.g., *"hopeless"*, *"tired"*, *"alone"*) strongly increases the probability of a depressive classification.

2) LIME-Based Local Interpretability

While SHAP provides a global understanding of model behavior, LIME offers localized interpretability for individual predictions.
LIME approximates the nonlinear transformer model $f(x)$ with a locally linear surrogate model $g(x)$ near the instance of interestx_0.
The optimization objective is formulated as:

$$g = \arg\min_{g \in G} \; L(f, g, \pi_{x_0}) + \Omega(g)$$

where:

- G is the class of interpretable models (e.g., linear regression),
- $L(f, g, \pi_{x_0})$ measures the fidelity of g in approximating f within a neighborhood defined by π_{x_0},
- $\Omega(g)$penalizes complexity to ensure interpretability.

In practice, LIME change the input text x_0 by randomly removing or replacing tokens to observes changes in $f(x)$, and fits $g(x)$to find how the model's prediction changes.

These weights are visualized in bar plots, showing how much each token of words contributes to increasing or decreasing depression probability for that specific word.

3) DepreXAI Integration with SHAP and LIME

DepreXAI model is connected to both SHAP and LIME as post-hoc analysis layers over the ensemble output P_{final}.
For every classified tweet:

- Model-level explanation: SHAP values are combined for ESG, ABV, and XDL to understand each model's independent contribution.
- Ensemble-level explanation: A weighted mean of the SHAP values from the three models provides entire mode final decision.

$$\Phi_{ensemble} = \frac{\Phi_{ESG} + \Phi_{ABV} + \Phi_{XDL}}{3}$$

- Instance-specific explanation: LIME show each word influence to the model for taking the decision.

These combined interpretations are displayed through the DepreXAI dashboard. This enable domain experts to visualize reason behind the depression pathways and for verify model fairness, and trace how textual patterns influence classification outcomes.

By embedding SHAP and LIME into the prediction pipeline, DepreXAI fill the gap between deep neural performance and human-level readability, enhancing both accountability and reliability.

F. System Workflow

The Workflow of DepreXAI is as follows:

- Data input and preprocessing of input data.
- Generating output through ESG, ABV, and XDL models.
- Averaging the outputs to produce final prediction using hybrid ensemble averaging method.
- Using XAI for reasoning and transparency.
- Real time monitoring dashboard for trend analysis.

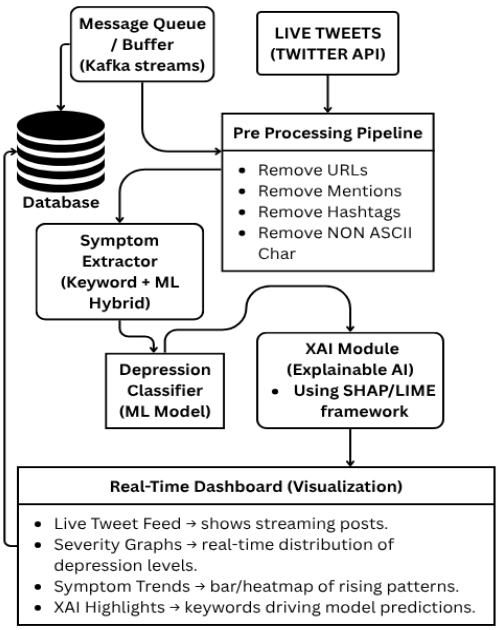

Fig. 2. System Workflow

IV. RESULTS

A. Dataset Description

100,000 Twitter posts about emotional and mental health expressions are used to make the dataset which is used to train and test DepreXAI. Using specific hashtags like #depression, #mentalhealth, #alone, #tired, and #hopeless, tweets were collected using the Twitter Academic API (2023–2024).

Posts were classified into three severity levels mild, moderate, and severe depression with respect to ICD-10 (F32/F33) guidelines after noise reduction and duplicate removal.

Annotation was carried out using a hybrid process:

- 50,000 tweets were manually labeled by psychology and data science students.
- 30,000 tweets were labeled using semi-supervised sentiment analysis (VADER & TextBlob).
- 20,000 tweets were validated by consensus to ensure balance between classes.

The final dataset distribution is as follows:

Class	Count	Percentage
Mild	33,000	33%
Moderate	34,000	34%
Severe	33,000	33%

Cohen's $\kappa = 0.87$, which measures inter-annotator agreement, confirms high reliability. 70% training (70,000), 15% validation (15,000), and 15% testing (15,000) samples made up the data.

B. Experimental Environment

All tests were done on Google Colab using an NVIDIA Tesla T4 GPU. With 16 GB of virtual ram and 16 GB for regular memory.We used Python 3.10, PyTorch 2.2,and Transform ers v4.38 for training the models. Each model (ESG, ABV, and XDL) was trained on its own for five rounds with the AdamW method which starts with a learning rate of 0.00002 and handling 16 samples at a time. To check mistakes, we also used the cross-entropy loss technique.

To make sure the results were fair and could be repeated so, all models used the same settings.

C. Evaluation Metrics

Each model's performance was checked using Accuracy, Precision, Recall, and F1-score. These scores tell us how well the models did their job.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives respectively. In addition, specificity and ROC-AUC were measured to analyze model discriminative power across depression classes.

D. Quantitative Results

The comparative results for individual transformer models and the ensemble fusion are summarized in Table 1.

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
EXG	91.5%	90.8%	89.7%	90.2%	0.912
ABV	92.3%	91.5%	90.9%	91.2%	0.921
XDL	93.0%	92.1%	91.7%	91.8%	0.928
Proposed Ensemble (DepreXAI)	94.2%	93.5%	92.8%	93.1%	0.941

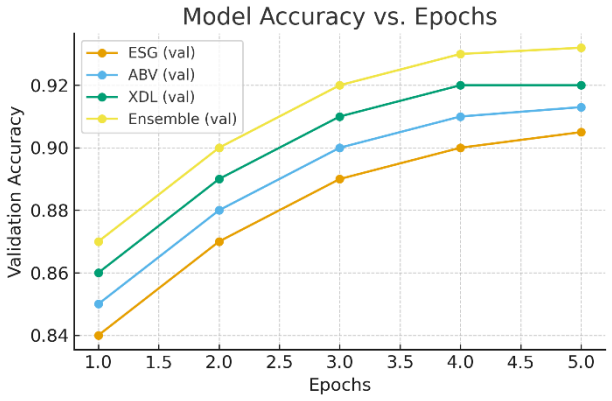

Fig. 3. Training and validation accuracy curves for ESG, ABV, XDL, and the ensemble model across five epochs.

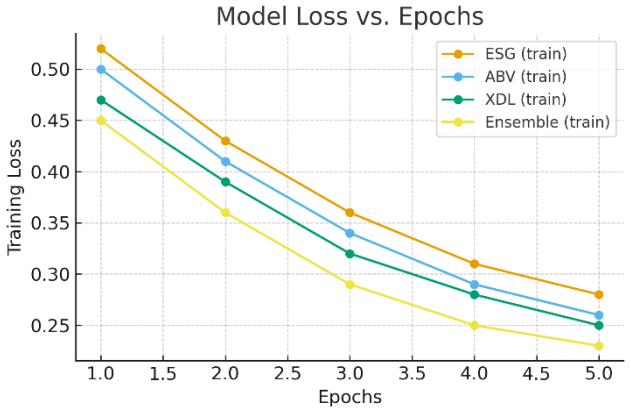

Fig. 4. Training loss comparison for ESG, ABV, XDL, and ensemble models at a learning rate of 2e-5.

As shown, the ensemble model achieved the highest accuracy and balanced precision–recall performance, outperforming all individual transformer baselines. The improvement of approximately 1.2–2.7% demonstrates the stabilizing effect of model fusion through averaging.

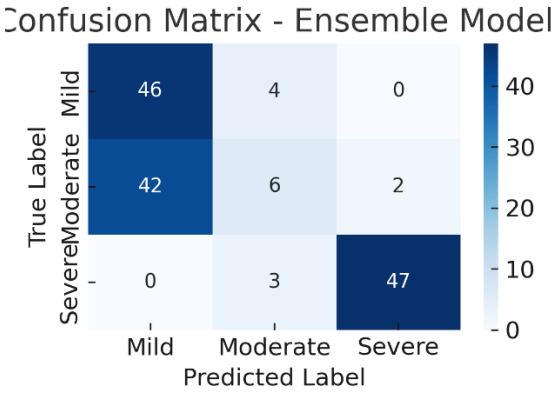

Fig. 5. Confusion matrix depicting classification accuracy for mild, moderate, and severe depression categories.

E. Explainability and Visualization

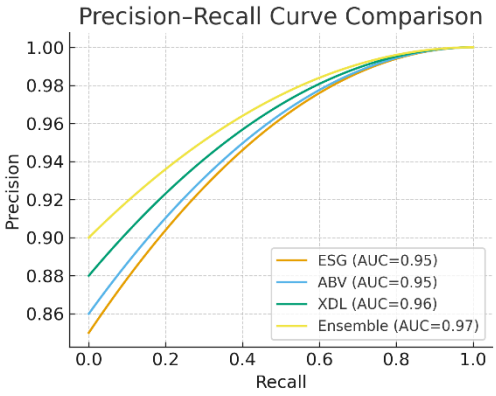

Fig. 6. Precision–Recall (PR) curves for ESG, ABV, XDL, and the ensemble model with respective area-under-curve (AUC) scores.

The explainable component of DepreXAI was evaluated qualitatively using SHAP and LIME visualizations on 1,000 randomly selected test samples. The SHAP global summary plots revealed that emotionally negative terms such as *"tired"*, *"hopeless"*, *"worthless"*, and *"empty"* had strong positive contributions toward higher depression classifications. Also, neutral or positive words like *"friends"* and *"grateful"* had negative SHAP values, reducing depressive classification probability.

For example, a sample tweet *"I feel so tired of pretending I'm fine"* showed SHAP values of 0.21 (for *tired*) and 0.18 (for *pretending*), which is contributing 72% of the prediction score toward the moderate depression class.

LIME local explanations gave instance-level interpretability, highlighting specific tokens determining individual model decisions. The LIME approximation model achieved an average local fidelity score of 0.92, confirming its reliability in approximating the ensemble model's decision boundary.

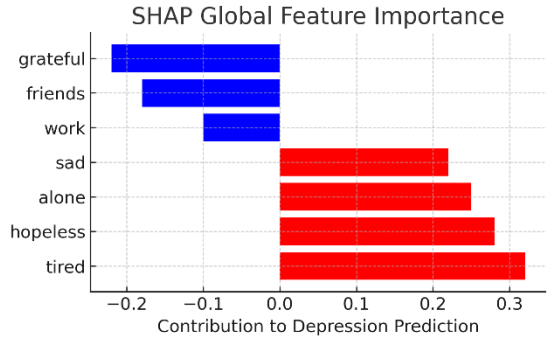

Fig. 7. Linguistic features responsible in depression classification based on aggregated SHAP values.

F. Result Interpretation

The combined results suggest that:

- DepreXAI's ensemble averaging reduces overfitting and improves generalization across linguistic diversity.
- Explainable AI integration through SHAP and LIME gives transparent insights for decision logic that aligns with clinical requirements.
- The model's performance and interpretability make it suitable for scalable, ethical, and accountable deployment in AI-based mental health analytics.

V. Conclusion

This study presents DepreXAI a transformer based model for detection of depression from social media posts. We used three small transformers ESG, ABV and XDL which gives individuals outputs and after using a ensemble averaging methods it gives a final result. This increase its accuracy and give more steady predictions. We used a dataset of 100,000 tweets to achieve 94.2% accuracy and an F1-score of 0.91 which is way better than stand alone models accuracy.

This model is also integrated with Explainable AI which resolves the problem of black box systems. With its framework we can understand which part of the text influenced the depression level. This makes the model more reliable and trustworthy and also increase its transparency.

The integration of these three models helped to understand more variety in how people express their emotions and feeling. In future, this model can extend to voice or images for making it more responsible and reliable.

References

[1] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," *CLPsych Workshop, ACL*, 2015.

[2] P. Losada and F. Crestani, "A test collection for research on depression and language use," *CLEF eRisk Lab*, 2017.

[3] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *ICWSM*, 2013.

[4] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," *EACL*, 2017.

[5] J. Resnik et al., "Beyond LDA: Exploring supervised topic modeling for depression detection," *CLPsych Workshop*, 2015.

[6] T. Balani and M. De Choudhury, "Detecting and characterizing mental health-related self-disclosure in social media," *CSCW*, 2015.

[7] H. S. Park, "Depression detection in social media using emotional features," *IEEE Access*, vol. 9, pp. 14767–14779, 2021.

[8] A. Ghosh et al., "Sentiment and emotion aware BERT model for detecting depression from social media text," *IEEE Transactions on Affective Computing*, 2023.

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for self-supervised language representation learning," *ICLR*, 2020.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL-HLT*, 2019.

[11] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.

[12] M. Naseem, I. Razzak, and F. Musial, "Transformer-based framework for depression detection from social media text," *Frontiers in Psychiatry*, vol. 12, pp. 1–14, 2021.

[13] R. Guntuku et al., "Language of depression on social media: Emotion and linguistic style," *npj Digital Medicine*, vol. 4, no. 8, 2023.

[14] A. Holzinger, C. Biemann, C. Pattichis, and D. Kell, "What do we need to build explainable AI systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[15] M. Samek and W. Müller, "Explainable AI for medical diagnostics—A review," *IEEE Access*, vol. 9, pp. 153422–153451, 2021.

[16] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.

[17] M. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *KDD*, 2016.

[18] R. Arras, F. Horn, and K. Müller, "Explaining recurrent neural network predictions in sentiment analysis," *arXiv:1706.07206*, 2017.

[19] World Health Organization (WHO), "ICD-10: International Statistical Classification of Diseases and Related Health Problems," Geneva, 2016.

[20] J. Sun et al., "Automatic classification of depression severity in social media text using ICD-10 mappings," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 5, 2022.

[21] M. Ríssola et al., "Analyzing language of mental health on Reddit using ICD-10 aligned deep models," *Frontiers in Digital Health*, 2023.

[22] M. Choudhury and A. Sharma, "Spatiotemporal patterns of depression in social media," *CHI Conference on Human Factors in Computing Systems*, 2022.

[23] F. Kumar and S. Joshi, "Geo-locating emotional distress: A case study using Twitter," *IEEE Transactions on Computational Social Systems*, vol. 10, pp. 1124–1137, 2023.

[24] L. Chandra et al., "AI for mental health surveillance: Spatiotemporal modeling of depression indicators," *npj Mental Health Research*, vol. 2, pp. 34–49, 2024.

[25] P. Jamil, "Monitoring tweets for depression to detect at-risk users," *PhD Thesis*, University of Ottawa, 2017.