ADScAI
Applied Data Science & Artificial Intelligence Symposium

Data-Driven Depression Prediction: Insights from a Machine Learning Challenge

1st Chamindi M.G.H.A
Dept. of CSE
University of Moratuwa
Colombo, Sri Lanka
hiruni.22@cse.mrt.ac.lk

2nd Uthayasanker Thayasivam
Dept. of CSE
University of Moratuwa
Colombo, Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Keywords*—Depression prediction, machine learning, XGBoost, feature engineering, mental health**

I. Introduction

Depression is a prevalent mental health condition that significantly impacts quality of life. Traditional diagnostic methods rely on clinical assessments and self-reported questionnaires, which can be subjective and time-consuming. The increasing availability of digital health data presents an opportunity for automated and data-driven depression screening using machine learning techniques.

While previous research has demonstrated the feasibility of AI-driven depression detection, challenges remain in feature selection, handling imbalanced datasets, and ensuring model interpretability. This study develops a machine-learning model using structured questionnaire data to predict depression, addressing these challenges and improving screening accuracy. The paper discusses relevant literature, dataset preprocessing, model selection, results, and future directions, concluding with insights and implications of our findings.

II. Literature Review

Machine learning has become an essential tool in the early detection of depression, with various models and techniques developed to identify individuals at risk. Early research primarily utilized statistical methods such as logistic regression and support vector machines (SVMs) applied to structured clinical and demographic data. While these models offer interpretability, they often struggle to capture the complex, nonlinear relationships present in mental health data. More advanced techniques, including deep learning models such as neural networks and transformers, have been used to analyze text-based data from social media posts, electronic health records, and chat logs to detect depressive symptoms.

Feature selection is a critical factor influencing predictive performance. Studies have examined diverse feature sets, including:

- **Clinical and demographic features**: Age, gender, socioeconomic status, and medical history.
- **Behavioral patterns**: Sleep activity, physical movement (via wearable devices), and smartphone usage.
- **Textual and sentiment analysis**: Evaluating social media posts, chat logs, and written diaries for depressive language patterns.

While structured clinical datasets provide reliable labels, they are often limited in size. To overcome this, researchers have explored large-scale self-reported surveys and passive data collection methods, which enhance scalability but pose challenges in data quality and reliability. Additionally, feature engineering techniques have been developed to incorporate behavioral and textual data, improving model accuracy. However, challenges remain in obtaining high-quality labeled datasets and ensuring generalizability across diverse populations.

Building on prior research, our study applies advanced machine learning techniques to structured questionnaire responses, leveraging XGBoost for improved predictive performance. By focusing on structured, interpretable data while maintaining high accuracy, this study contributes to the growing body of research in AI-driven mental health analytics.

III. Materials and Methods

A. Dataset and Preprocessing

The dataset used in this study originates from the "Predicting Depression: Machine Learning Challenge" hosted on Kaggle. It consists of demographic, behavioral, and psychological attributes collected from participants. The target variable indicates whether an individual is likely to be depressed, making this a binary classification problem.

Before training the model, extensive data preprocessing was performed. Missing values were handled using different techniques: categorical variables were filled with the most frequent value, while numerical variables were imputed using K-Nearest Neighbors (KNN) imputation. Additionally, categorical features were encoded to ensure compatibility with machine learning algorithms. Some variables were transformed

https://doi.org/10.31705/ADScAI.2025.12

to better capture their relationships with depression, such as grouping certain professions into broader categories and normalizing numerical attributes.

B. Model Selection and Training

Given the dataset's nature and the need for a robust predictive model, we selected XGBoost as the primary algorithm. XGBoost is a gradient-boosting framework known for its efficiency, scalability, and high predictive performance. It has been widely used in machine learning competitions and real-world applications due to its ability to handle missing values and feature interactions effectively.

To optimize performance, we conducted hyperparameter tuning using a grid search approach. Parameters such as the number of estimators, learning rate, and maximum tree depth were fine-tuned to enhance the model's predictive ability. We employed 10-fold cross-validation to ensure the generalization of the model and mitigate overfitting.

XGBoost achieved 93.85 percent accuracy, outperforming logistic regression and random forests, with an F1-score of 0.8283 and ROC-AUC of 0.9735, ensuring robust classification.

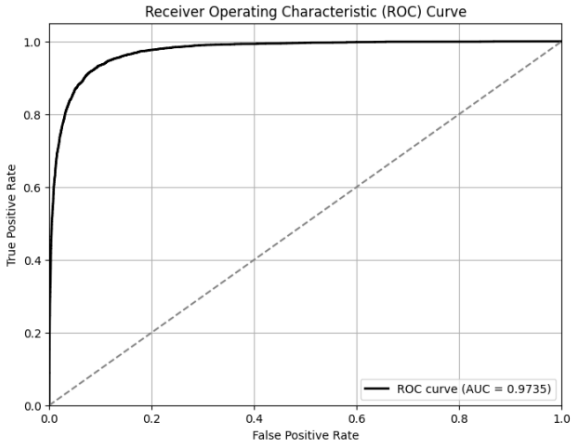

Fig. 1. ROC curve for the XGBoost model, demonstrating an AUC of 0.9735.

IV. Results and Discussion

The XGBoost model demonstrated strong performance in depression prediction, achieving 93.85 percent accuracy, an F1-score of 0.8283, and an ROC-AUC score of 0.9735, confirming its reliability and discriminatory power. A comparative analysis with logistic regression and random forests showed that XGBoost outperformed these models in predictive precision and generalization, reinforcing its suitability for depression classification.

Hyperparameter tuning and cross-validation played a crucial role in improving model stability, mitigating overfitting, and improving overall reliability. However, limitations include the reliance on self-reported questionnaire data, which is subject to recall bias and social desirability bias, which can lead to inaccurate labels. Furthermore, cultural and demographic variations in depression symptoms can affect the generalizability of the model. Future research should explore passive data collection methods (e.g., wearable devices and social media analysis) to complement self-reported data and improve prediction accuracy.

V. Ethical Considerations

Ethical considerations are paramount in AI-driven mental health applications. Ensuring data privacy requires robust anonymization techniques and adherence to regulations such as GDPR and HIPAA. Additionally, model interpretability remains a challenge, as black-box models like XGBoost can obscure decision-making processes, making clinical integration difficult. Future work should focus on explainable AI (XAI) methods to improve transparency.

In addition, fairness concerns must be addressed to prevent bias against underrepresented demographic groups. Adopting ethical AI frameworks, such as those proposed by the WHO and FAT-ML, can help guide responsible deployment in real-world settings. It is also critical to ensure that AI-driven mental health tools complement rather than replace traditional clinical assessments, maintaining the involvement of healthcare professionals in the diagnostic process.

VI. Conclusion

This study explored the application of machine learning for depression prediction using data from the "Predicting Depression: Machine Learning Challenge." Through comprehensive data preprocessing and model selection, we developed an XGBoost model that demonstrated strong predictive performance in identifying individuals at risk of depression.

While the model outperformed traditional machine learning techniques, certain limitations remain, including the reliance on self-reported data and the need for more diverse feature sets. Future research could enhance these models by integrating additional data sources, such as social media behavior or physiological signals, to improve accuracy and generalization.

This research contributes to the growing field of AI-driven mental health applications by demonstrating the feasibility of machine learning in depression prediction. Ethical considerations and data privacy remain critical concerns, highlighting the need for responsible AI deployment in healthcare settings. Future work should also explore interpretability methods to ensure that clinicians and mental health professionals can effectively utilize these models in practice.

References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 22, pp. 785–794, August 2016.

[2] U. Madububambachu, A. Ukpebor, and U. Ihezue, "Machine Learning Techniques to Predict Mental Health Diagnoses: A Systematic Literature Review," Clinical Practice and Epidemiology in Mental Health, vol. 20, pp. e17450179315688, July 2024.

[3] A. Sharma and W. J. M. I. Verbeke, "Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset," Frontiers in Big Data, vol. 3, Apr. 2020.