Comparative Analysis of Machine Learning Models for Optimized Detection and Personalization in Mental Health

Ritik Samanta
PGDM Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
ritik.samanta2325p@ves.ac.in

Prof. Nikita Ramrakhiani
PGDM Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
nikita.ramrakhiani@ves.ac.in

Mrudula Geddam
PGDM Business Analytics
Vivekanand Education Society's
Business School
Mumbai, India
geddam.mrudula2325p@ves.ac.in

Abstract—**The growing concern for mental health urgently calls for intervention in this often-ignored domain of general health. Stigma and barriers to access have created large treatment gaps in mental health disorders. Major depressive and anxiety disorders, each affecting more than a tenth of the population, confer the highest global burden in India according to the World Health Organization. The Post pandemic era has seen an increase in the levels of stress, social isolation, and subsequently the prevalence of mental health disorders. The study investigates advanced, data-driven methods to enhance the detection and intervention of mental health conditions. This paper compares various predictive models, such as Logistic Regression, Decision Tree, and Random Forest, in terms of their effectiveness concerning predictions related to mental health. Such a comparison becomes important since it helps draw an effective, systematic comparison of the performances of each model with respect to accuracy, precision, and overall effectiveness. Such machine learning models will then be compared with traditional diagnostic methodologies to identify which gives the most valid and timely identification of a mental health disorder. The comparative study of different machine learning models is important, as it gives not only various strengths and weaknesses of those models but also enough guidance on choosing an appropriate approach in conducting diagnostics and interventions for mental health and will enable not only the optimization of predictive analytics in mental health care but also more effective elaboration of better-targeted treatment strategies. Data visualization with tools such as Power BI and Tableau presents an important approach in this investigation, where increased functionality for interactive, actionable presentation for complex data is being enabled by giving proper assessment of risk about mental health and making effective intervention planning through dynamic visual dashboards. Real-time monitoring of mental health trends can identify how effectively interventions can work and, thus, optimize the treatment strategy. The study explores the status of mental health and its association with selected demographic features such as gender, state of residence, living status of family members, and educational background. Lifestyle variables that were taken into consideration are average sleep duration, income, and exercise, which were analyzed in relation to psychological outcomes. This would underscore the strengths and limitations of each model, namely Logistic Regression, Decision Tree, and Random Forest, in predicting mental health outcomes and, therefore, help guide further development of more focused and effective interventions. The paper foresees that the integration of advanced machine learning techniques with data visualization will remarkably upgrade diagnosis and management in mental health for better prevention and quality of care. This is an important step toward better mental health management in India and across the globe.**

Keywords—Mental health, Machine learning, Data visualization, Predictive modeling, Logistic regression, Decision tree, Random Forest, Gradient boosting machines.

I. Introduction

Mental health problems are common worldwide including changes in mood, personality, inability to cope with daily problems or stress, withdrawal from friends and activities, and so on. A mental health problem is defined as an impairment in a person's cognition, emotional control, or behavior patterns that have clinical significance and is often linked to distress or functional impairment (Moein Razavi, et. al). Increases in the occurrence and global burden of mental illness have made the prevention and treatment of mental disorders a public health priority [90, 91, 204, 207]. In India, recent studies suggest that approximately 15-20% of the population suffers from some form of mental health disorder. Mental health conditions cause a great deal of distress or impairment; depression alone will affect 11% of the world's population (Ariel Rosenfeld, et. al). However, it's crucial to remember that these are estimates, and the actual prevalence may be higher due to underreporting and stigma associated with mental health issues in India. Some cases cause minor hindrances in someone's day-to-day life whereas others can result in someone taking their own life. There is no certain explanation as to what someone ends up doing as a result of their deteriorating mental health [3]. This makes mental health disorders a serious issue that needs to be dealt with swiftly. Mental health refers to a state of well-being in which an individual realizes his or her own abilities to cope with the normal stresses of life, work productively and fruitfully, and is able to make useful contributions to his or her society and community. (Maria Tamoor, et. al)

The ability to detect early warning signs of depression, continuously and as effortlessly as possible, by extending current assessment methods could have a significant impact in mitigating or addressing depression and its related negative consequences. Various efforts have been made to identify risk factors of suicide thoughts and behaviors, and to predict the probability of future suicide attempts.

Although a few high-performance models were reported in studies where the cohorts were enriched for cases, prediction accuracy was limited when applied to a general population (Le Zheng, et. al).

In this research we aimed at comparing and identifying an advanced machine-learning technology to incorporate the relations between different risk factors for mental health diseases using a dataset containing a comprehensive profile of patients. Machine learning studies generally differ from traditional research in two ways. The first is a focus on prediction rather than inference. The second is a shift towards model flexibility, with the ability to handle large numbers of predictors simultaneously (Adam M. Chekroud, et. al). There are very few widely used or validated biomarkers in mental health, leading to a heavy reliance on patient- and clinician-derived questionnaire data as well as interpretation of new signals such as digital phenotyping (Ariel Rosenfeld, et. al). Adrian B. R. Shatte et.al identified four main application domains in the literature, including: (i) detection and diagnosis; (ii) prognosis, treatment and support; (iii) public health, and; (iv) research and clinical administration and most of the ML techniques used included support vector machines, decision trees, neural networks, latent Dirichlet allocation, and clustering.

As AI techniques continue to be refined and improved, it will be possible to help mental health practitioners re-define mental illnesses and identify these illnesses at an earlier or prodromal stage when interventions may be more effective, and personalize treatments based on an individual's unique characteristics (Sarah Graham, et. al). Yena Lee et. al highlighted that Predictive models integrating multiple data types performed better when compared to models with single lower-dimension data types. For our research, the questionnaire was designed with less reliance on clinical questions and information but more on psychosocial factors, daily routine, sleep cycle, lifestyle habits, sufferance from chronic diseases, physical activity, etc. The dataset, which comprised 200,000 records from Indian states between 2017 and 2021, profiled patients across physiological health, psychosocial factors, and lifestyle habits. It included 16 socio-demographic indicators and lifestyle parameters.

This data was essential for developing targeted mental health care strategies, enhancing the understanding of how socio-demographic factors influenced mental well-being and improving patient outcomes.

All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceeding. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. Research Methodology

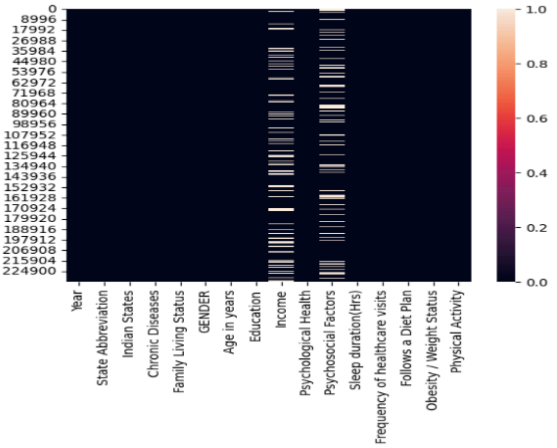
This section outlines the methodology employed in this study to detect mental health from user profiles. The methodology encompasses data collection, preprocessing, feature extraction, and the training and comparison of machine learning algorithms logistic regression, decision tree, random forest and XGBoost gradient boosting. A comparative analysis was done to get a clear understanding of the performance of the four algorithms. The objective is to determine the important predictive variables and evaluate how well machine learning models do in predicting mental health disorder early detection.

A. *Dataset*

Data was collected from multiple secondary sources defining the profile of patients in four sections: physiological health, psychosocial factors, standard of living and lifestyle habits. The dataset spans between 2017 and 2021, with approximately 200,000 records from Indian states. Information on 16 socio-demographic indicators and lifestyle parameters like chronic disease, physical activity, and diet parameters would be outlined. Such indicators would hence be important for formulating customized mental health care strategies.

B. *Data Preprocessing*

A preliminary data cleaning revealed that 15% of the data was missing and mainly pertained to the socio-economic variables, specifically income and psychological health. For income, we used median imputation, while for psychological health, we made use of a custom imputation. This way, we ensured that our imputation will not affect too much the general flow and distribution of the dataset.



-

Fig 1: Distribution of Data Before Imputation

In Figure I, we have demonstrated data distribution before imputation. The missing values through imputation appropriately filled in without driving much of the central tendencies or spread of variables, so the structure of the dataset was preserved in the most reasonable way. The study's hypothesis provides a framework for analyzing the relationship between demographic and lifestyle variables and psychological health outcomes.

***Hypothesis 1*:** Differences in psychological health are associated with variations in demographic factors such as gender, state of residence, family living status, and educational background.

***Hypothesis 2*:** Lifestyle factors such as average sleep duration, income levels, and physical activity significantly influence psychological health outcomes.

***Hypothesis 3*:** Machine learning-based predictive models offer earlier and more accurate detection of mental health disorders compared to conventional diagnostic methods.

C. *Model Training and Evaluation*

We trained and tested four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and XGBoost Gradient Boosting. In each model, accuracy, precision, and potential overfitting were considered for their relative strengths and weaknesses to be plotted.

III. Exploratory Data Analysis (EDA)

Exploratory data analysis, conducted for this research, provides a comprehensive overview of demographic and geographic variables and the relationship between lifestyle factors and psychological health, thus detailing insight into the data.

A. *Gender-Based Analysis of Psychological Health*

Figure.2, illustrates the distribution of psychological health across genders. The analysis reveals that mental health issues are prevalent in both males and females, with slight differences: 23.75% of males report "Good" psychological health compared to 23.58% of females. This near parity suggests that both genders experience mental health challenges; however, the marginally better outcomes for males highlight the importance of gender-specific mental health services. Despite the small difference, it underscores the necessity for tailored interventions that address the unique needs of both genders.

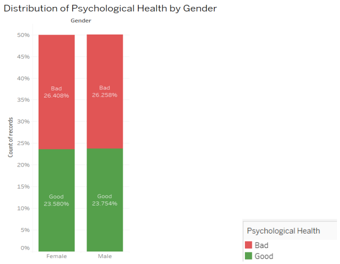


Fig 2: Psychological Health by Gender Distribution

B. *Geographical Analysis of Mental Health by State*

Figure.3, presents a heatmap illustrating the geographical distribution of psychological health across Indian states. The analysis indicates that states such as Punjab and Maharashtra face more significant mental health challenges, while regions like Lakshadweep and the Andaman and Nicobar Islands report fewer issues. These disparities may stem from either underreporting or genuine differences in the demand for mental health care. The map visualization effectively highlights this contrast, with graded colors representing states with varying levels of mental health concerns. These findings underscore the need for region-specific mental health policies that better address the unique needs of each area.

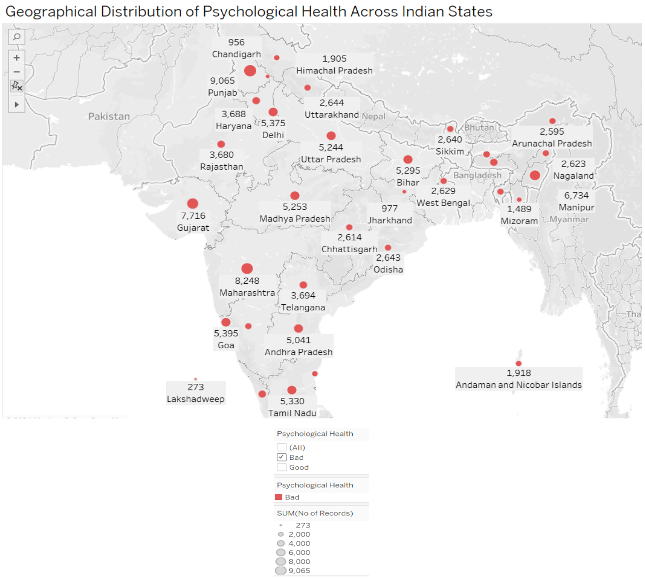


Fig 3: Geographical Distribution of Psychological Health

C. *Family Living Status and Psychological Health*

Figure.4, illustrates the relationship between family living arrangements and psychological health. The data reveals that individuals living alone have a 21.31% probability of reporting "Bad" psychological health, while those residing with roommates report a slightly lower rate of 15.48%. In contrast, individuals in joint or nuclear families exhibit significantly lower rates of poor psychological health, at approximately 7.95%. This analysis, as depicted in Figure 4, highlights the importance of social support systems in mitigating psychological distress. It suggests that family structures, which promote interaction and support, may not only be neutral but potentially protective against mental health issues.

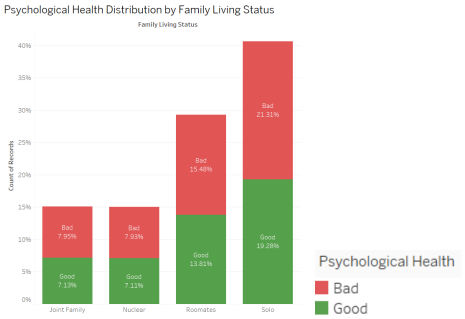


Fig 4: Psychological Health by Family Living Status

D. *Sleep Duration and Psychological Health*

Figure.5, illustrates the relationship between average sleep duration and psychological health outcomes. The data analysis indicates that individuals experiencing poorer mental health tend to sleep longer. For instance, those categorized as having "Bad" mental health logged a total of

946,760 hours of sleep, compared to 850,250 hours for those classified as having "Good" mental health. These findings, as shown in Figure 5, suggest that extended sleep may serve as a coping mechanism for individuals facing psychological challenges, acting as a form of psychological adaptation.

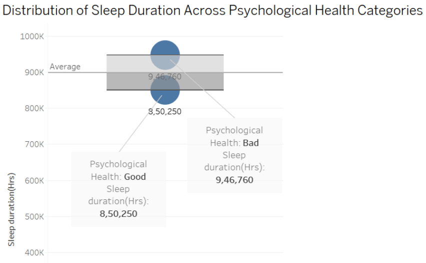


Fig 5: Distribution of Total Sleep Hours by Psychological Health Categories

E. *Income Distribution and Psychological Health*

Figure 6 depicts the income distribution before data cleaning, revealing a right-skewed distribution. Initially, most individuals are situated in lower-income brackets, while a smaller proportion falls into higher brackets. This skewness, observed during data preprocessing, suggests a potential link between income distribution and mental health, particularly due to income-related stressors. Figure 6 emphasizes the necessity for adjusting income variability to obtain unbiased insights into its relationship with psychological health.

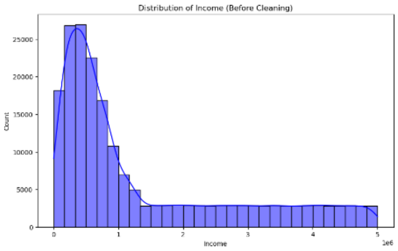


Fig 6: Income Distribution Before Data Cleaning

F. *Skewness and Kurtosis Analysis of Key Variables*

Figure 7 presents the skewness and kurtosis analysis of key variables, highlighting their statistical properties. Income shows a skewness of 1.46 and a kurtosis of 0.87, while physical activity displays even greater skewness (3.37) and kurtosis (9.35). In contrast, psychological health appears to be approximately normally distributed. The histogram in Figure 7 illustrates these differences, indicating that both income and physical activity may exhibit erratic distributions and heavy tails. This suggests that failing to account for these properties in the analysis could negatively impact the accuracy of the models.

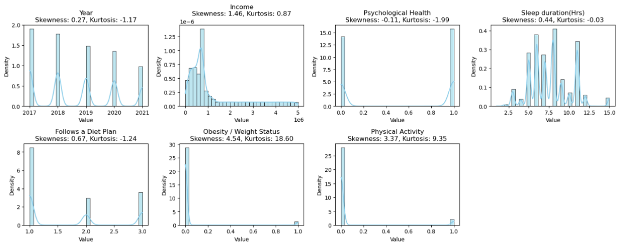


Fig 7: Skewness and Kurtosis of Key Variables

The EDA uncovers essential insights into how demographic and lifestyle factors shape mental health outcomes. Through visualizations, the analysis deepens comprehension and engagement with the data. These findings underscore the importance of targeted interventions that address gender differences, regional disparities, family support structures, lifestyle habits, and socioeconomic conditions. This comprehensive analysis sets the stage for further research and practical applications in mental health care.

IV. RESULTS

In our study on the machine learning-based prediction of psychological health, we tested a number of algorithms with astonishing results. For the logistic regression model, an accuracy of 66.38% meant that it had a good ability in selecting cases of poor psychological health although it failed to be precise for the "Good" class. The decision tree model reflected the accuracy at 66.00%, high aptitudes towards detecting complex patterns but together with overfitting issues at the same time.

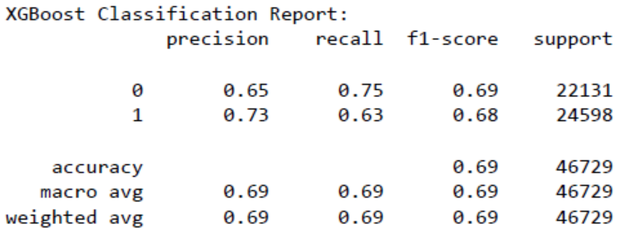
Improving the accuracy of the model to 67.39%, the random forest model was successful in reducing overfitting while emphasizing important predictors, such as income and education. The best performance accuracy occurred with the XGBoost, with a 68.63% metric.

XGBoost was chosen for its powerful gradient boosting capabilities, which sequentially build trees to correct the errors of previous models, enhancing predictive accuracy, especially in complex datasets. This approach is more effective than traditional methods like Decision Trees and Random Forests, which do not leverage this error-correcting mechanism. Additionally, Logistic Regression struggles with non-linear relationships without significant feature engineering. The learning rate hyperparameter in XGBoost allows for fine-tuning the impact of each tree, effectively balancing high accuracy with the risk of overfitting, a challenge that other models may not address as effectively.

XGBoost also incorporates both L1 and L2 regularization techniques to mitigate overfitting, a feature not typically found in Decision Trees and Random Forests by default. While Logistic Regression offers some level of regularization, it often falls short in managing complex datasets compared to the robust regularization provided by XGBoost's trees. Furthermore, XGBoost excels in scalability and efficiency, utilizing advanced tree pruning strategies such as “max depth” and “min child weight” to maintain model complexity without unnecessary growth.

This prevents overfitting while ensuring that critical features are preserved.

The model's ability to leverage parallel processing and distribute computing makes it particularly well-suited for handling large datasets, such as those encountered in mental health research. These combined features make XGBoost an ideal choice for this analysis, ensuring reliable predictions and efficient processing.



```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.75      0.69     22131
           1       0.73      0.63      0.68     24598

    accuracy                           0.69     46729
   macro avg       0.69      0.69      0.69     46729
weighted avg       0.69      0.69      0.69     46729
```

Fig 8: XGBoost Classification Report Analysis

As we see in Figure 8 XGBoost demonstrated a balanced precision and recall, achieving values of 0.73 for precision and 0.63 for recall in class 1, effectively managing both true positives and true negatives. In contrast, Logistic Regression and Decision Trees tended to skew towards one class, increasing the risk of false positives or negatives.
The F1-scores for XGBoost were also impressive, at 0.69 for class 0 and 0.68 for class 1, indicating a strong reliability in identifying both positive and negative cases.

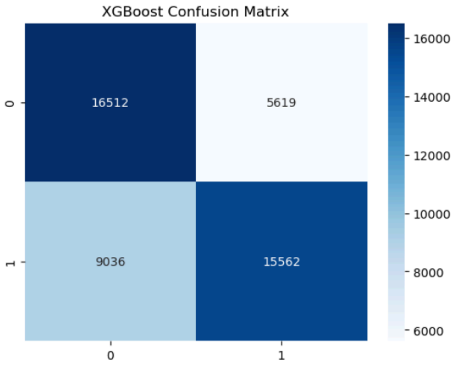


Fig 9: XGBoost Classification Report Analysis

The confusion matrix in Figure 9 further highlighted XGBoost's performance, showing a higher count of correctly predicted samples across both classes compared to other models. This lower misclassification rate can be attributed to XGBoost's iterative corrections and feature prioritization, which were not as effectively replicated by Logistic Regression or Decision Trees.

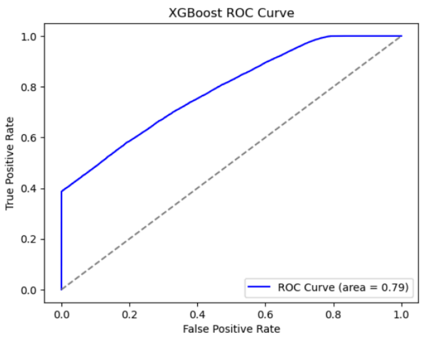
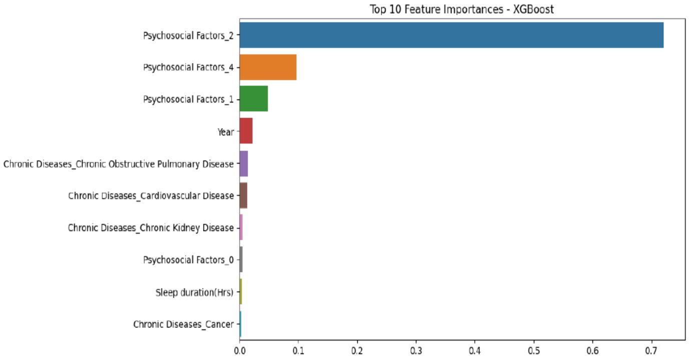


Fig 10 : ROC Curve

Figure 10 shows ROC curve and AUC score, XGBoost achieved a higher AUC of 0.79, reflecting a superior ability to separate positive and negative classes. The steep rise in the ROC curve indicated a significant increase in true positives with minimal false positives, showcasing XGBoost's refined boundary-setting capability. Additionally, the balanced shape of the ROC curve demonstrated that XGBoost effectively distinguishes between classes across various thresholds, unlike other models that showed rapid declines or plateaus in true positive rates.



The feature importance analysis from the XGBoost model reveals that psychosocial factors play a crucial role in predicting mental health outcomes. The top features identified, such as "Psychosocial Factors_2," "Psychosocial Factors_4," and "Psychosocial Factors_1," significantly influence the model's predictions, underscoring the importance of addressing these factors in mental health assessments and interventions. This analysis not only highlights the predictive power of these features but also suggests that targeted interventions focusing on psychosocial elements could enhance mental health outcomes. Furthermore, the ability of XGBoost to calculate importance weights based on the reduction of error across trees allows for a nuanced understanding of feature interactions, providing valuable insights for healthcare providers and policymakers. By prioritizing these influential factors, resources can be allocated more effectively, ultimately improving strategies aimed at promoting mental well-being.

XGBoost’s ROC curve reveals a robust ability to separate classes with high sensitivity, especially in the initial range of thresholds, allowing accurate detection of positive cases with minimized false positives. Its feature engineering approach, particularly the gradient-weighted feature importance and sparse data handling, allowed the model to leverage nuanced interactions between predictors, reducing noise from less significant variables. Regularization controls ensured the model remained generalizable and avoided overfitting, an advantage the other models lacked.

The results indicate that even advanced machine learning techniques can enhance predictions about psychological health and direct more focused interventions in terms of mental health.

V. LIMITATION

According to the model the four models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—had exhibited significant limitations when applied to complex mental health data. Logistic Regression had failed to capture non-linear relationships, assuming a linear connection between features and the target variable, which had led to overlooked insights. Decision Trees had been prone to overfitting, particularly with noisy data, resulting in low generalizability and inconsistent predictions due to their sensitivity to small changes in the dataset. While Random Forest had addressed overfitting by averaging predictions from multiple trees, it had sacrificed interpretability, making it difficult to understand how specific mental health features contributed to outcomes. XGBoost, despite its accuracy, had faced challenges related to computational intensity and the potential for overfitting if regularization parameters had not been properly tuned. These limitations had highlighted the necessity for more advanced modeling techniques to effectively analyze mental health indicators and enhance predictive accuracy.

VI. FUTURE SCOPE

For future work, accuracy could be enhanced by exploring deep learning models like Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) networks, which will capture complex and temporal patterns within mental health data. Hybrid and stacked ensemble models, such as combinations of LightGBM, CatBoost, and XGBoost, will further leverage strengths across models to boost predictive power and reduce overfitting. Automated feature engineering techniques, like Feature Tools, along with dimensionality reduction methods (e.g., PCA), will identify high-impact features, improving data quality and model efficiency.

References

[1] S. Patel and C. Pietrabissa (2021), "Introduction to Digital Mental Health," in Digital Mental Health, 1st ed., M. Faure, Ed. Elsevier. Available: https://doi.org/10.1016/B978-0-12-820203-6.00001-1.

[2] M. R. Iniesta-Sepúlveda (2021), "Machine Learning in Mental Health: A Scoping Review of Methods and Applications," Psychological Medicine, Cambridge University Press, vol. 49, pp. 381-393. Available: https://doi.org/10.1017/S0033291719003150.

[3] D. Althoff (2019), "Social Media and Mental Health Disorders: An Overview," Current Psychiatry Reports, Springer, vol. 21, pp. 1-9. Available: https://doi.org/10.1007/S11920-019-1094-0.

[4] H. Chien (2024), "Digital Mental Health Interventions for Depression and Anxiety," JMIR Mental Health, vol. 11, pp. e53714. Available: https://doi.org/10.2196/53714.

[5] A. C. Halverson and J. K. Martin (2018), "Sleep and Mental Health Correlates in Depression," Journal of Affective Disorders, Elsevier, vol. 250, pp. 32-41. Available: https://doi.org/10.1016/j.jad.2018.03.004.

[6] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen (2019), "Mental Health Monitoring with Multimodal Sensing and Machine Learning: A Survey," Pervasive and Mobile Computing, vol. 51, pp. 1-25. Available: https://doi.org/10.1016/j.pmcj.2019.01.008.

[7] B. G. Bokolo and Q. Liu (2023), "Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques," Electronics, vol. 12, p. 4396. Available: https://doi.org/10.3390/electronics12214396.