A Comparative Analysis of Data Mining Techniques for Mental Health Prediction Among University Students in Bangladesh

Abstract—**In this paper, we propose a data-driven approach to identify and predict mental health issues such as stress, anxiety, and depression among students using data mining techniques. University students in Bangladesh face increasing mental health challenges, which are often unrecognized until severe consequences occur. Using a structured data set of 485 records obtained from a publicly available survey. The data set comprises 39 features in various domains, including demographic attributes, academic stress indicators, emotional and behavioral symptoms (such as hopelessness, nervousness, and suicidal thoughts), and self-reported coping mechanisms. The preprocessing steps included data cleaning, removing outliers, normalization, and handling of class imbalance using SMOTE. Applied both Classification models, Logistic Regression, Support Vector Machines, Random Forests, and advanced deep learning architectures such as CNNs and LSTMs. Feature selection techniques, used Pearson correlation, information gain, and gain ratio, were employed to optimize model input. The results indicate that Gradient Boosting and Decision Trees achieved near-perfect accuracy, while deep learning models showed substantial performance after balancing. This research demonstrates the methods of data driving to detect mental health risks and provides a foundation for developing early intervention systems tailored to the regional academic context. This research demonstrates the viability of data-driven methods in detecting mental health risks and provides a foundation for developing early intervention systems tailored to the regional academic context.**

Index Terms—**Keywords: Mental Health Prediction, Student Dataset, Data Mining, Machine Learning, Deep Learning, SMOTE, Feature Engineering, Bangladesh Universities.**

I. Introduction

In June 2023, a second-year student from the Department of Public Administration at the University of Rajshahi tragically ended his life. In his suicide note, he expressed overwhelming feelings of loneliness that had consumed him. This heartbreaking accident underscores the profound mental health challenges faced by university students in Bangladesh. According to Bangladesh Post, this case is part of a disturbing trend of rising mental health issues and suicides among the student population (Bangladesh Post, 2023). Such tragedies are not isolated. In 2023, at least 98 university students in Bangladesh died by suicide, contributing to a total of 513 student suicides nationwide. A recent study published in the Dhaka Tribune revealed that 75.85% of university level students in Bangladesh experience mental health challenges, driven by post-pandemic academic pressure, financial stress, and the absence of effective mental health support systems (Dhaka Tribune, 2023).

There are suicide cases in Dhaka city due to the lack of a supportive environment for students to study and grow, according to experts. Despite these alarming statistics, mental health remains a stigmatized and neglected issue in many Bangladeshi universities. Most institutions lack trained counselors, mental health infrastructure, and formal policies to support student well-being. To address this growing crisis, it is essential to integrate mental health education into academic curricula, establish accessible psychological counseling services, and foster an open and supportive environment where students can seek help without fear of stigma or judgment. Therefore, there is an urgent need not only to provide accessible psychological support but also to implement early identification systems such as screening tools or mental wellness programs that can help determine whether a student is mentally ill or at risk, before it's too late.

II. Related Works

Recently, many AI, data mining, and machine learning approaches have been used to detect mental health diseases and disorders. In one study [1], the diagnosis of mental health disorders has been the subject of research. Models such as Support Vector Machines (SVM), Random Forest, and neural networks have been shown to efficiently assess clinical and behavioral data for early diagnosis, with a primary focus on child depression. The Random Forest model among these has attained up to 95 percent accuracy and 99 percent precision.

In this paper, [2] reduced a 90-question SCL-90-R survey to 28 questions while diagnosing 10 mental disorders with 89 percent precision using the network pattern recognition algorithm (NEPAR). The algorithm was trained using Lasso logistic regression, Ridge logistic regression, random forest, and SVM. Lasso Logistic Regression with NEPAR-P yielded the best accuracy; however, it lacked Mult disorder coverage and consideration of ethical AI.

In this paper, [3] proposed a mental health assessment system using an improved Decision Tree and ANN algorithm, achieving higher precision and system efficiency. It utilized data mining techniques on student mental health datasets to optimize prediction and classification. However, existing studies often relied on traditional questionnaires, lacked adaptability, and did not leverage real-time big data analytics.

In this paper, [4] predicted student mental health using an online survey of 109 responses, applying Logistic Regression with hyperparameter tuning for best results. The study found strong correlations between depression, anxiety, social connectedness, and academic performance, especially among early-year students. But most related models lacked focus on regional contexts like Odisha and didn't explore personalized intervention strategies.

This paper [5] uses Logistic Regression, K-NN, Decision Tree, Random Forest, and Stacking to determine a mental health illness. Here, every single method is applied to a survey dataset and does not achieve the best accuracy. When applying an ensemble method like Stacking, get the best accuracy of 81.75 percent. This study mainly focused on improved determination through ensemble methods. Here, do not apply any deep learning model, explore explainable AI, and do not use multisource data integration.

In this paper, [6] applied Naive Bayes, Random Forest, SVM, KNN, and Decision Tree to identify mental health issues using survey data. Among all of the models, Naive Bayes achieved a high accuracy of 65.91 percent. In the paper, we show a strong correlation between daily life and depression. This study highlights only Bangladeshi students, uses a real-world, locally collected dataset, and develops this specific region for mental health prediction. Here, do not use any deep learning and ensemble methods, a small data set, and a lack of external validation.

This paper [7] works on data mining and machine learning to predict Generalized Anxiety Disorder and tests three algorithms like Naïve Bayes, Random Forest, and J48, applying the Shapley value for selecting the best features. The results showed that only important features improved more accuracy of determining key symptoms like suicidal thoughts and tiredness. Here uses a small data set (180 women) and lacks data diversity. In future work, it needs a big data set and truly different machine learning models, and implement a real health setting system to detect GAD early.

In addition, a recent survey [8] shows that advanced deep learning models such as LSTM and BERT have shown promising potential to analyze social media posts, identifying signs of depression and anxiety with accuracies up to 93 percent. There are still obstacles to overcome despite these encouraging advancements, such as problems with data quality, moral dilemmas, and the requirement for interpretable models.

Recently,[9], the diagnosis of mental health disorders has been the subject of research. It has been shown that models like Support Vector Machines (SVM), Random Forest, and neural networks can efficiently assess clinical and behavioral data for early diagnosis, with a primary focus on child depression. Up to 95 percent accuracy and 99 percent precision have been attained by the Random Forest model among these.

This paper works to predict anxiety, depression stress labels using some unique variants which is not been used in other papers and applies 5 different classifications like Decision Tree, Random Forest Tree, Naïve Bias, Support Vector Machine, and K-Nearest Neighbor. Among all of them, Naïve Bias has the highest accuracy, 85.5 percent, but overall, the Random Forest is best based on F1 score. Its limitation was class imbalance. In the future, it will use balance and large data [10].

In this paper,[11] works with machine learning models to predict mental health problems in adolescents using parent-reported data. Like Random Forest Tree, Support Vector Machine, Neural Network, XGBoost, and Logistic Regression are used to predict. Among them, Random Forest gives the best accuracy of 73.9 percent. Its main limitation was parent-reported data and class imbalance. In the future, it will use parent surveys, balance, and large data.

This paper works only XGBoost algorithm to predict mental illness onset using medical check-up data and wearable data from 4612 people and achieves an accuracy of 71.2 percent. Sleep-activated related features are more important than other features. Here, used small number of people's mental illness cases, missed and neglected symptoms, only Fitbit, wearable data, and medical tests did not come at the same time. In the future, it will work on positive cases and build a personal time-series model for better accuracy[12].

This study focuses on mental health conditions and mental illness prediction and academic features among university students using a survey dataset (39 features). This dataset covers all over students, like academic performance, emotions, behavioral symptoms, and self-reports, including nervousness, hopelessness, sleeping problems, and suicidal thoughts. The model applies machine learning and deep learning models like Logistic Regression, Random Forest, SVM, Naïve Bayes, KNN, Decision Tree, Gradient Boosting (GBM), ANN, CNN 1D, RNN(LSTM). It also notices class imbalance with SMOTE and ensures more reliable prediction across all severity levels, and ethical AI principles like fairness, privacy protection, and responsible model use

III. Methodology

A. Data Collection

The data [13] used in this study were obtained from a publicly accessible online repository and represent secondary data originally gathered through a structured mental health survey. The survey targeted university students in Bangladesh to understand their mental health status in the context of academic and social pressures. The data set was not collected by the authors but was recovered for academic research purposes according to ethical data usage guidelines. All responses were anonymized, with no personally identifiable information present. The dataset comprises responses from Bangladeshi university students, totaling 485 individual entries and encompassing 22 features. These features represent a diverse mix of thematic categories and data types relevant to the study of student mental health. Demographic variables such as age, gender, and university or department affiliation provide an essential context to analyze mental health outcomes across different student groups. Academic stressors are also well represented, with attributes that include exam pressure, fear of failure, workload, and peer comparison, factors commonly cited in the literature as contributing to student distress. The data set also includes key indicators of psychological well-being, such as self-reported experiences of depression, anxiety, suicidal thoughts, and the likelihood of seeking professional

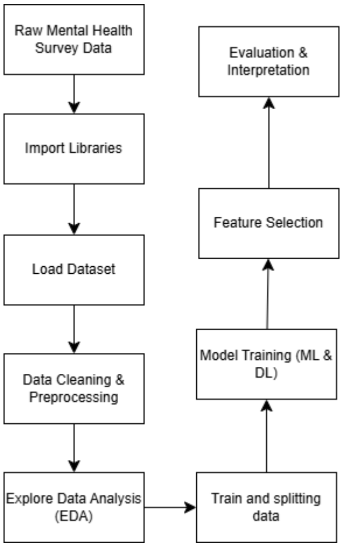

Raw Mental Health Survey Data
Import Libraries
Load Dataset
Data Cleaning & Preprocessing
Explore Data Analysis (EDA)
Train and splitting data
Model Training (ML & DL)
Feature Selection
Evaluation & Interpretation

Fig. 1. Work Flow

help. In addition, the data captures information on support systems and coping strategies, such as communication with friends or family, consultation with psychologists, and engagement with motivational content. The dataset integrates a variety of data types, including categorical responses, for example, "Yes" or "No", "Male" or "Female", ordinal values from Likert-type scales, and numerical values such as age. This comprehensive and multidimensional data structure is particularly well-suited for exploratory data analysis, statistical correlation assessments, and the development of predictive machine learning models aimed at identifying students at elevated mental health risk.

B. *Workflow*

The workflow illustrated in Figure-1 outlines the process applied to an existing mental health dataset. Subsequently, machine learning and deep learning techniques were used to assess their effectiveness in classifying mental health conditions, including stress, anxiety, and depression. The study used a publicly available online mental health survey dataset, representing secondary data. The data set includes responses on mental health conditions, workplace policies, demographics, and treatment-seeking behavior.

The analysis environment was initialized by importing essential Python libraries such as Pandas and NumPy for data manipulation, Matplotlib, Seaborn, and Plotly for visualizations, Scikit-learn for model training and evaluation. The raw data was loaded into a pandas DataFrame for subsequent processing. At this stage, the structure and contents of the dataset were explored to understand the attribute types and distributions. To enhance the quality and usability of the dataset, a series of data cleaning and preprocessing tasks were undertaken. Initially, non-essential columns such as comments, state, and Timestamp were removed, as they did not contribute meaningful information for analytical or predictive purposes. The dataset's column names were standardized by removing special characters and formatting inconsistencies to ensure uniformity and compatibility with the analytical tools used. In addition, outlier handling was performed on the age column by filtering out values below 18 and above 100, thereby retaining only plausible responses within a reasonable age range. A critical part of preprocessing involved normalizing categorical entries particularly gender. Due to the presence of inconsistent and varied textual responses, all gender entries were consolidated into three standardized categories: Male, Female, and Other. These steps collectively improved the dataset's reliability for analysis and machine learning modeling. After preprocessing, Exploratory Data Analysis (EDA) was conducted to better understand the characteristics of the dataset and uncover underlying patterns. Demographic variables such as age and gender were explored through visualizations, including histograms and bar plots, to assess their their distributions. Additionally, mental health-related attributes were analyzed to identify key trends. This included examining treatment-seeking behavior across genders, the perceived interference of mental health issues with job performance, awareness of mental health support benefits provided by employers, and the anticipated consequences of disclosing mental health conditions at work. These analyses offered valuable insights into the relationships among various factors and guided the selection of features for model development. Following EDA, the dataset was divided into training and testing subsets. This step was essential to evaluate model performance in a controlled manner and to ensure that the models could generalize to new data. Typically, a standard split of 80% training, 20% testing was applied, enabling robust evaluation of the trained models. This separation also helped prevent overfitting by validating the model on unseen data. With the training set prepared, machine learning algorithms were employed to predict mental health treatment-seeking behavior. Algorithms such as Logistic Regression and Random Forest were used for classification tasks. These models were trained on the selected features using the Scikit-learn .fit() method, enabling them to learn patterns and relationships within the training data. The training process aimed to optimize the models' ability to distinguish between individuals likely or unlikely to seek mental health treatment based on their responses. After initial training, feature selection was carried out to improve model efficiency and performance. This involved analyzing the correlation between features and the target variable, as well as assessing the relative importance of each feature using model-based techniques such as feature importance scores derived from the Random Forest model. Features that showed little to no relevance were excluded, helping to reduce noise and computational overhead during model training and evaluation. The final models were evaluated using the testing subset to determine their predictive performance. Standard classification metrics such as accuracy, precision, recall, and F1-score were used to measure the effectiveness of each model. These metrics provided a comprehensive view of the model's ability to classify instances and handle imbalanced class distributions correctly. The results were interpreted to assess the feasibil-

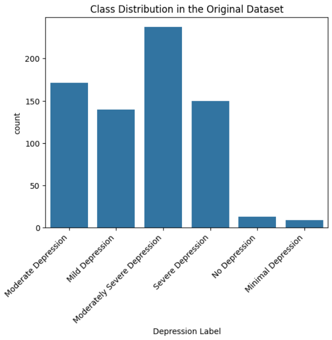


Fig. 2. Original dataset

ity and practical implications of using machine learning for predicting mental health. Ultimately, the evaluation helped validate the potential of data-driven approaches in understanding behavioral health trends and informed the study's broader conclusions.

IV. Result and Analysis

Experiments were conducted with both traditional and deep learning models to evaluate performance on the dataset. Models of traditional machine learning include Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (DT), and ensemble methods like Random Forest and Gradient Boosting. Furthermore, deep learning models like ANN, CNN, and RNN were included. Accuracy, precision, recall, and F1 score are common classification metrics that were used to assess each model after it was trained on the preprocessed dataset. These metrics not only measure how many predictions were correct, but also identify the students with mental health issues. Initially, it is observed that the method GBM and the DT model performed extremely well, achieving perfect scores. Logistic Regression and SVM also produced good results with accuracies over 90%. However, models like RNN and Naive Bayes performed poorly, most likely as a result of the dataset's imbalance, which favored one class. Low recall and F1 scores resulted from this imbalance, which also hampered their ability to identify cases of the minority class. For balancing the dataset, SMOTE was applied. Significant gains were seen in all categories after retraining the models on the balanced sample. The accuracy of models that initially performed poorly, such as RNN, improved significantly (from 47.92% to 86.11%), other models, such as SVM and KNN, also showed improvements in recall and F1 score. However, the performance of CNN slightly dropped, possibly due to its sensitivity to the structure of the input data, which may have been affected by synthetic samples. To enhance model performance and reduce data dimensionality, three different feature selection methods were applied. Firstly, remove features with high intercorrelation using Pearson correlation (threshold ¿ 0.8) to minimize multicollinearity. Then, the information gain was computed using mutual information to evaluate how strongly each feature was associated with the target class. Finally, the Gain Ratio was

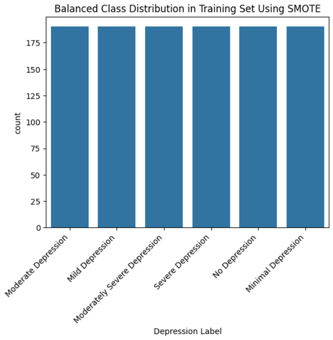


Fig. 3. Balanced Class distribution

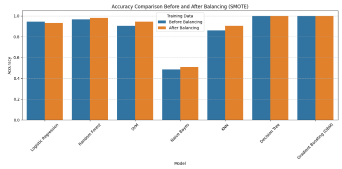


Fig. 4. Ml model accuracy Before and After Balancing

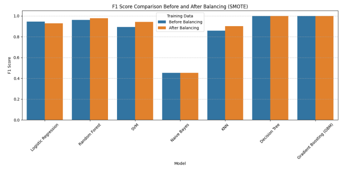


Fig. 5. Ml model F1 score Before and After Balancing

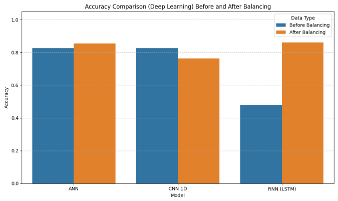


Fig. 6. Advanced model accuracy score Before and After Balancing

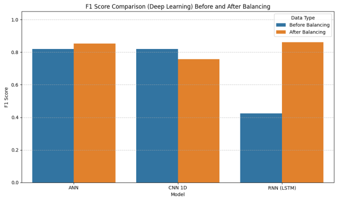


Fig. 7. Advanced model F1 score Before and After Balancing

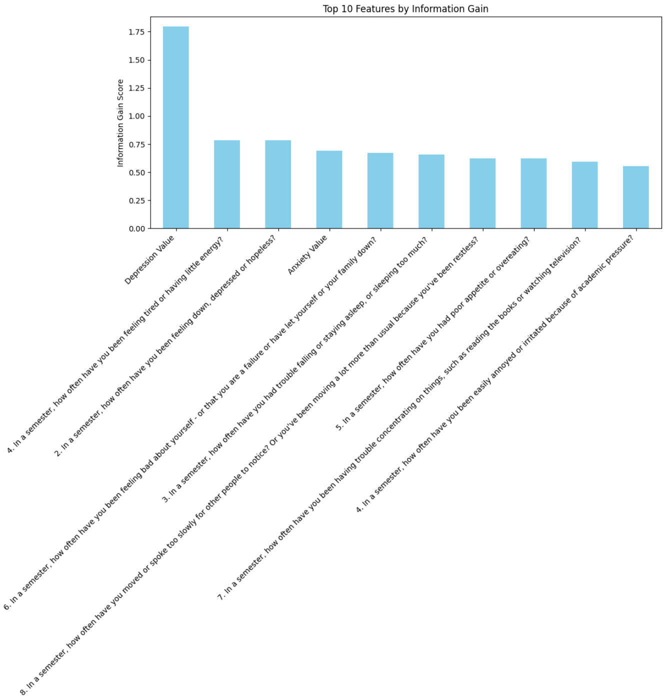
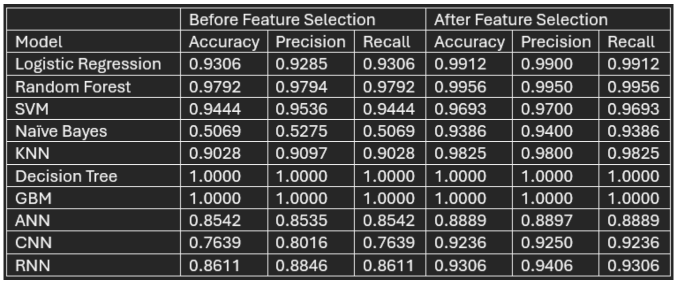


Fig. 8. Feature selection



	Before Feature Selection			After Feature Selection		
Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Logistic Regression	0.9306	0.9285	0.9306	0.9912	0.9900	0.9912
Random Forest	0.9792	0.9794	0.9792	0.9956	0.9950	0.9956
SVM	0.9444	0.9536	0.9444	0.9693	0.9700	0.9693
Naïve Bayes	0.5069	0.5275	0.5069	0.9386	0.9400	0.9386
KNN	0.9028	0.9097	0.9028	0.9825	0.9800	0.9825
Decision Tree	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
ANN	0.8542	0.8535	0.8542	0.8889	0.8897	0.8889
CNN	0.7639	0.8016	0.7639	0.9236	0.9250	0.9236
RNN	0.8611	0.8846	0.8611	0.9306	0.9406	0.9306

Fig. 9. Before and after feature selection

calculated by dividing the Information Gain by each feature's intrinsic entropy to avoid overvaluing features with many distinct values. After comparing and visualizing the results of these techniques, using the information gained, the most informative features were selected for the final model training.

With the selected feature, both the ML models and deep learning models performed well. The precision of the logistic regression increased from 0.9306 to 0.9912, and the accuracy of the Random Forest increased from 0.9792 to 0.9956. Naive Bayes gives a better improvement from 0.5069 to 0.9386, which indicates a high sensitivity to irrelevant features. Deep learning models also show better performance with selected features.

V. Conclusion

In this paper, it demonstrates the potential of data mining and machine learning techniques in effectively predicting mental health problems such as stress, anxiety, and depression among university students in Bangladesh. By utilizing a comprehensive dataset of 485 student records and applying a variety of models including both traditional machine learning algorithms and deep learning architectures, the research highlights the strengths and limitations of different approaches. Gradient Boosting and Decision Trees delivered the highest accuracy, while class imbalance, initially a significant challenge, was successfully mitigated using the SMOTE technique. The application of feature selection methods further improved model performance by enhancing precision and reducing noise. Although the study relies on self-reported survey data, which may not be as reliable as clinical assessments, the findings lay a strong foundation for early intervention systems tailored to academic settings in Bangladesh. Future work could integrate clinical data and behavioral analytics in real time to create more robust, ethically sound, and actionable mental health monitoring tools for educational institutions.

References

[1] S. Tutun, M. E. Johnson, A. Ahmed, A. Albizri, S. Irgil, I. Yesilkaya, E. N. Ucar, T. Sengun, and A. Harfouche, "An ai-based decision support system for predicting mental health disorders," *Information Systems Frontiers*, vol. 25, no. 3, pp. 1261–1276, 2023.

[2] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification algorithms based mental health prediction using data mining," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1174–1178.

[3] B. Sahu, J. Kedia, V. Ranjan, B. P. Mahaptra, and S. Dehuri, "Mental health prediction in students using data mining techniques," *The Open Bioinformatics Journal*, vol. 16, no. 1, 2023.

[4] M. Luo, "Research on students' mental health based on data mining algorithms," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 1382559, 2021.

[5] K. Vaishnavi, U. N. Kamath, B. A. Rao, and N. S. Reddy, "Predicting mental health illness using machine learning algorithms," in *Journal of Physics: Conference Series*, vol. 2161, no. 1. IOP Publishing, 2022, p. 012021.

[6] M. N. Hossain, N. Fahad, R. Ahmed, A. Sen, M. S. Al Huda, and M. I. Hossen, "Preventing student's mental health problems with the help of data mining," *International Journal of Computing*, vol. 23, no. 1, pp. 101–108, 2024.

[7] N. Jothi, W. Husain, and N. A. Rashid, "Predicting generalized anxiety disorder among women using shapley value," *Journal of infection and public health*, vol. 14, no. 1, pp. 103–108, 2021.

[8] M. Garg, "Mental health analysis in social media posts: a survey," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 1819–1842, 2023.

[9] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS one*, vol. 16, no. 12, p. e0261131, 2021.

[10] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020.

[11] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," *PloS one*, vol. 15, no. 4, p. e0230389, 2020.

[12] T. Saito, H. Suzuki, and A. Kishi, "Predictive modeling of mental illness onset using wearable devices and medical examination data: machine learning approach," *Frontiers in digital health*, vol. 4, p. 861808, 2022.

[13] M. S. Islam, "Bangladeshi university students mental health dataset," https://doi.org/10.6084/m9.figshare.25347691.v1, 2024, figshare Dataset.