International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENHANCING BIPOLAR DISORDER DETECTION USING UNSUPERVISED CLUSTERING-DRIVEN OPTIMIZED XGBOOST: A GAUSSIAN MIXTURE MODEL AND SMOTE-BASED HYBRID APPROACH

[1]Santosh Rani, [2]Dr. Neeraj Mangla

[1]Research Scholar

Computer Science and Engineering Department

MMEC, Maharishi Markandeshwar (Deemed to be University) Mullana, Ambala, India

Santoshmehta1423@gmail.com

[2]Professor

Computer Science and Engineering Department

MMEC, Maharishi Markandeshwar (Deemed to be University) Mullana, Ambala, India

neerajmangla@mmumullana.org

Abstract:

Bipolar Disorder (BD) is a serious mental health condition that is often difficult to diagnose correctly due to overlapping symptoms and imbalanced clinical data. Early and accurate diagnosis can help patients receive timely treatment and support. In this study, we developed a hybrid machine learning approach that integrates Gaussian Mixture Model (GMM)**,** Synthetic Minority Over-sampling Technique (SMOTE), and an optimized XGBoost classifier to improve the detection of BD.

First, GMM was applied to uncover hidden clusters of patients based on medical and psychometric features, generating informative cluster assignments as additional features. Next, SMOTE was used to address the imbalance between bipolar and non-bipolar cases by generating synthetic samples for the minority class. Finally, an optimized XGBoost model was trained on the enriched and balanced dataset, leveraging both original clinical features and GMM-derived cluster representations for improved predictive power.

The GMM-SMOTE-XGBoost hybrid model achieved an accuracy of 93.27% on an independent test set. Comparative analysis against conventional classifiers such as Logistic Regression(LR), Decision Tree(DT) and Support Vector Machine(SVM), and shows that this hybrid approach consistently outperformed all baselines, particularly in sensitivity and overall diagnostic reliability.

This work demonstrates the potential of combining unsupervised clustering, synthetic data balancing, and optimized gradient boosting to enhance psychiatric disorder classification. Such hybrid frameworks could serve as valuable clinical decision-support tools, improving diagnostic accuracy and robustness in mental health diagnostics.

Keywords: Bipolar Disorder Prediction, Gaussian Mixture Model, Synthetic Minority Over-sampling Technique, Hybrid Machine Learning Model.

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. Introduction

BD is a severe and chronic mental health condition characterized by distinct and often unpredictable shifts between episodes of mania, hypomania, depression, and mixed states. This inherent dynamism presents considerable challenges for clinicians in achieving timely and accurate diagnoses, implementing effective long-term monitoring strategies, and ultimately delivering personalized treatment plans that can adapt to the fluctuating nature of the illness. The significant impact of BD on an individual's quality of life, functional capacity, and overall well-being underscores the critical need for advanced analytical tools that can shed light on its complex progression. The episodic and recurrent nature of BD means that patients experience highly individualized trajectories through various mood states. Current clinical approaches often rely on retrospective patient reports or infrequent assessments, which can miss crucial transitional periods or early warning signs of an impending episode. Without a clear understanding of these individual trajectories and the ability to predict future mood state transitions, interventions may be reactive rather than proactive, leading to prolonged suffering, increased hospitalization rates, and suboptimal treatment outcomes. Therefore, there is a pressing need for robust predictive models that can capture the longitudinal patterns of BD and forecast mood shifts.

However, most existing models are either limited by severe class imbalance or fail to capture the hidden heterogeneity within BD populations. This research proposes a novel hybrid framework that combines the unsupervised clustering capability of the GMM with the predictive power of an optimized XGBoost classifier**.** GMM-derived cluster labels are incorporated as auxiliary features to enhance the representation of latent patient subgroups, while the SMOTE is applied to mitigate class imbalance and improve the model's ability to detect minority bipolar cases.

By integrating clustering-based feature enrichment, synthetic data balancing, and an optimized gradient boosting model, this approach aims to achieve superior accuracy, recall, and overall diagnostic robustness. The hybrid GMM-SMOTE-XGBoost model is rigorously evaluated and compared against traditional classifiers including LR, DT, SVM. Results demonstrate that the proposed framework significantly improves both sensitivity and specificity, highlighting its potential as a reliable clinical decision-support tool for BD detection.

2. Related Works

Qiu et al. [1] have proposed a system consists of two layers: a local motion direction detection layer and an unsupervised global motion direction detection layer. For local motion detection, adopted the Local Motion Detection Neuron (LMDN) model, which detects motion in eight different directions. The outputs of these neurons serve as inputs to the global motion direction detection layer, which employs a GMM for unsupervised clustering.

Pal and Paul [2] have investigated how much K-means and GMM suffers from uneven class distribution in data. Later experiment on benchmark imbalanced datasets with different imbalance ratio and over sampled datasets using SMOTE has been carried out for proposed approach. For each case cluster forest has been used as an attribute selection technique. Efficacy of the proposed Boosted GMM approach compared to standard clustering approaches like K means and GMM is exhibited from empirical analysis.

Zhang and Yang [3] have represented an improved SMOTE algorithm with appropriate sampling space for high dimensional data. Then the GMM-based synthetic sampling approach

Received: August 05, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

will be proposed. Afterward, an adaptive optimization method is proposed in our study for the hyperparameters of sampling process.

Pan et al. [4] have aimed to compare the accuracy of SVM and GMM in the detection of manic state of BDs of single patients and multiple patients. Methods: 21 hospitalized BD patients (14 females, average age 34.5±15.3) were recruited after admission. Spontaneous speech was collected through a preloaded smartphone.

Hassantabar et al. [5] have proposed a framework called MHDeep that utilizes commercially available WMSs and efficient DNN models to diagnose three important mental health disorders: schizoaffective, major depressive, and bipolar. MHDeep uses eight different categories of data obtained from sensors integrated in a smartwatch and smartphone. These categories include various physiological signals and additional information on motion patterns and environmental variables related to the wearer. MHDeep eliminates the need for manual feature engineering by directly operating on the data streams obtained from participants. Because the amount of data is limited, MHDeep uses a synthetic data generation module to augment real data with synthetic data drawn from the same probability distribution.

Jyothi and Verma [6] have proposed structure is also helpful in designing a computer-based tool for recognizing stress and mental health predictions. This research examines the different transformations like depression to Mania, from depression to BD, from BD to hypomania, and to identify pre-bipolar depression: LR, SVM and another machine. Learning Regression and k-nearest-neighbour analysis are used to analyze risk variables and output accuracy.

Singh et al. [7] have proposed an ideal solution to identify the sickness in the person by checking with the recorded dataset. The most preferred SVM, DT Classifier is used for this purpose. The initial goal of the DT is to create training ideal which is used to forecast the target variable class. The parameters considered here are anxiety disorder, depression disorder and the stress. Random Forest (RF) algorithm is applied to predict the illness in the people. The result obtained is to have accurate prediction level compared to the existing model.

Sivagnanam and Visalakshi [8] have introduces a novel Heterogeneous Ensemble Machine Learning (HEML) approach designed to detect BD, a significant healthcare challenge that demands precise and prompt diagnosis for effective treatment. The HEML method integrates multiple machines learning models, incorporating various physiological, behavioral, and contextual data from patients. By using a comprehensive feature selection technique, relevant features are extracted from each data source and utilized to train individual classifiers for detecting mental disorders. The classifiers include Adaboost, Decision Tree, K-nearest neighbors, Multilayer Perceptron, Random Forest, Relevance Vector Machine, and XGB, with Logistic Regression serving as the meta-model. This ensemble of classifiers enhances overall performance by capturing a wider range of characteristics related to mental disorders.

WU et al. [9] have collected heterogeneous digital phenotype data from 84 individuals with BD and 11 healthy controls. Five-fold cross-validation was employed for evaluation. The experimental results revealed that the Lasso and ElasticNet regression models were the most effective in predicting rating scale scores, and heterogeneous data performed better than homogeneous data, with a mean absolute error of 1.36 and 0.55 for HAM-D and YMRS, respectively; this margin of error meets medical requirements.

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Pan et al. [10] have aimed to compare the accuracy of SVM and GMM in the detection of manic state of BDs of single patients and multiple patients. SVM provided an appropriate tool for detecting manic state for single patients, whereas GMM worked better for multiple patients' manic state detection. Both of them could help doctors and patients for better diagnosis and mood state monitoring in different situations.

Li et al. [11] have compared the ISMOTE algorithm with seven mainstream oversampling algorithms, using three classifiers on thirteen public datasets from the KEEL, UCI, and Kaggle databases. Comparative analysis of 2D and 3D scatter plots revealed that ISMOTE yields more realistic data distributions. Experimental results demonstrated relative improvements in classifier performance, with F1-score, G-mean, and AUC increasing by 13.07%, 16.55%, and 7.94%, respectively. Furthermore, ISMOTE's parameter adaptability enables its application to multi-class imbalanced datasets.

Huth et al. [12] have performed between-group comparisons using linear mixed effects models for all three risk assessment tools. Additionally, we aimed to differentiate the risk groups using a linear support vector machine. We found no significant volume differences between the risk groups for all limbic structures during the main analysis.

Saha et al. [13] have proposed an ensemble of hybrid model-based techniques that aims to build a strong detection model that considers many psychological and sociodemographic characteristics of an individual to detect whether a person is depressed. Support vector machines (SVM) and multilayer perceptrons (MLP) are the two fundamental methods used to construct the suggested ensemble approach. The hybrid DeprMVM served as a meta-learner. In this study, the hybrid DeprMVM is a level-1 learner, whereas the SVM and MLP networks are level-0 learners. After the classifiers are trained and tested at level 0, their outputs are based on both the independent and dependent variables in the new data set that was used to train the meta-classifier. The training data class imbalance was reduced by applying the synthetic minority oversampling technique (SMOTE) and cluster sampling together, which improved the accuracy for detecting depression.

Kosolwattana et al. [14] have proposed a novel self-inspected adaptive SMOTE (SASMOTE) model that leverages an adaptive nearest neighborhood selection algorithm to identify the "visible" nearest neighbors, which are used to generate samples likely to fall into the minority class. To further enhance the quality of the generated samples, an uncertainty elimination via self-inspection approach is introduced in the proposed SASMOTE model. Its objective is to filter out the generated samples that are highly uncertain and inseparable with the majority class. The effectiveness of the proposed algorithm is compared with existing SMOTE-based algorithms and demonstrated through two real-world case studies in healthcare, including risk gene discovery and fatal congenital heart disease prediction.

Ogunseye et al. [15] have presented research responds to increased mental illness conditions worldwide and the need for efficient mental health care (MHC) through machine learning (ML) implementations. The datasets employed in this investigation belong to a Kaggle repository named "Mental Health Tech Survey." The surveys for the years 2014 and 2016 were downloaded and aggregated. The prediction results for bagging, stacking, LR, KNN, tree class, NN, RF, and Adaboost yielded 75.93%, 75.93%, 79.89%, 90.42%, 80.69%, 89.95%, 81.22%, and 81.75% respectively. The AdaBoost ML model performed data cleaning and prediction on the datasets, reaching an accuracy of 81.75%, which is good enough for decision-making. The results were further used with other ML models such as Random Forest (RF), K-Nearest Neighbor (KNN), bagging, and a few others, with reported accuracy ranging from 81.22% to

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

75.93% which is good enough for decision making. Out of all the models used for predicting mental health treatment outcomes, AdaBoost has the highest accuracy.

Balakrishna et al. [16] have aimed to enhance Major Depressive Disorder (MDD) prediction and diagnosis using hybrid machine learning methods, focusing on EEG data alongside clinical and demographic information. Employing various algorithms like CatBoost, Random Forest, XG Boost, XGB Random Forest, SVM with a linear kernel, and logistic regression with Elasticnet regularization, the study found that CatBoost achieved the highest accuracy of 93.1% in MDD prediction and diagnosis, surpassing other models. Additionally, the ensemble model combining XGBoost and Random Forest showed strong performance in ROC analysis, effectively discriminating between individuals with and without MDD.

3. Methodology

This study proposes a hybrid machine learning framework for the classification of BD by integrating Gaussian Mixture Model (GMM) clustering with an optimized XGBoost classifier**,** supported by class balancing through the Synthetic Minority Over-sampling Technique (SMOTE)**.** The methodology is systematically organized into four key stages: data pre-processing**,** feature enrichment through GMM-derived cluster assignments**,** synthetic data augmentation using SMOTE**,** and final classification with optimized XGBoost followed by performance evaluation. The overall workflow of the proposed GMM-SMOTE-XGBoost model is illustrated in Figure 1.

3.1. Dataset and Pre-Processing

Effective data pre-processing is fundamental to the performance of machine learning models, particularly in healthcare datasets that often suffer from missing values, outliers, and inconsistent scales. The BD dataset used in this study comprises 7500 samples and 54 clinical, diagnostic, and psychometric features.

3.1.1 Missing Value Handling and Normalization

Missing values were addressed using mean imputation, ensuring that no samples were lost during the modeling process. To standardize feature scales and eliminate the influence of outliers, Min-Max normalization was applied to rescale all continuous variables to the range [0,1]. The normalized score for each attribute is computed using Equation (1):

$$P_i = \frac{A_i - \mu_A}{\sigma_A} \qquad (1)$$

Where:

- P_i = Normalized value for feature *i*
- A_i = Original attribute value for feature *i*
- μ_A = Mean of attribute A
- σ_A = Standard deviation of attribute A

3.2. Feature Selection

To enhance model interpretability and reduce the risk of overfitting, SelectKBest feature selection with the ANOVA F-test was employed. This method ranks features based on their

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

statistical significance with respect to the target variable. The information gain guiding the feature selection is defined in Equation (2):

$$H(D) = -\sum_{i=1}^{k} p_i log_2 p_i \quad (2)$$

Where:

- p_i = Proportion of instances belonging to class i in dataset D
- k = Number of distinct classes

The entropy of the dataset, which quantifies uncertainty, is calculated using Equation (3):

$$H(D) = -\sum_{c=1}^{C} P(c) log P(c) \quad (3)$$

Where:

- $H(D)$ = Entropy of the dataset D
- C = Total number of classes
- $P(c)$ = Probability of class cc in the dataset

The distribution of information after partitioning on attribute D is computed using Equation (4):

$$H(D|F) = \sum_{k=1}^{K} \frac{|D_k|}{|D|} \times H(D_k) \quad (4)$$

3.3 Clustering and Feature Augmentation

To capture latent data patterns, the pre-processed data was clustered using a GMM. The GMM assigns each sample a probabilistic cluster label, which is appended to the original feature set, enriching the representation of complex relationships in the data. Simultaneously, SMOTE was applied to generate synthetic minority class instances, effectively addressing the class imbalance inherent in the dataset. The outputs of GMM clustering and SMOTE augmentation were combined, creating an expanded feature set for classification.

3.4 Classification and Evaluation

The augmented dataset was used to train a hybrid model (GMM+SMOTE+Optimized-XGBoost). The model was optimized to predict the presence or absence of BD based on the enriched feature representation. Model performance was rigorously evaluated using key metrics including: Accuracy, Precision, Recall (Sensitivity), F1 Score, ROC-AUC Score and Confusion Matrix. Validation was conducted on an independent test set to ensure generalizability. The proposed hybrid model demonstrated superior classification performance.

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

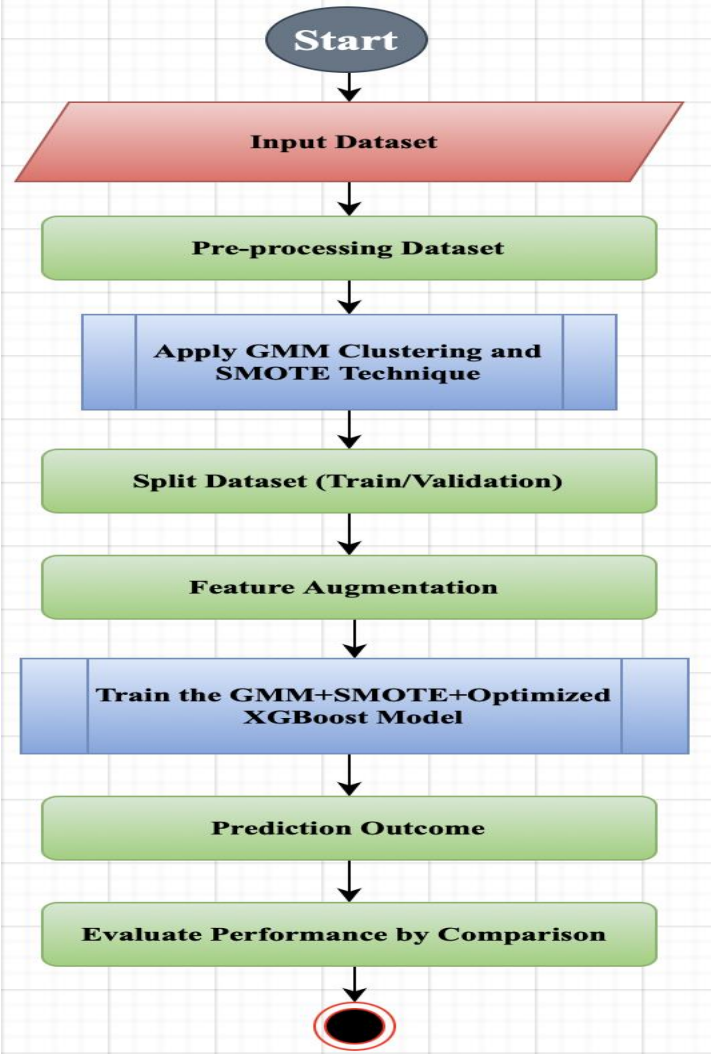

Figure 1. Overall Activity Diagram of the Model

Figure 1 illustrates the sequential workflow of the proposed hybrid classification model for BD detection**,** integrating GMM clustering with an optimized XGBoost classifier**,** along with SMOTE for class balancing**.** The process follows these key stages: data pre-processing, latent cluster extraction using GMM, synthetic minority oversampling with SMOTE, and final classification using optimized XGBoost.

The process begins with the initiation of the system, clinical, diagnostic, and psychometric data related to BD are provided as the input dataset and pre-processing is performed, including: Missing value imputation, Min-Max normalization (scaling features to [0,1]) and Encoding categorical variables.

The pre-processed data is divided into training and testing sets to enable model development and evaluation. The SMOTE is applied to the training set to address class imbalance by generating synthetic examples of the minority class. Simultaneously, a GMM is applied to uncover latent patterns in the data through unsupervised clustering. The cluster labels are used as additional informative features. The augmented dataset is used to train

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

a GMM+SMOTE+Optimized-XGBoost model, which leverages both the synthetic balanced data and the cluster information for improved classification accuracy. The trained model generates predictions for the test set, determining the likelihood of BD presence. Model performance is evaluated using key metrics such as: Accuracy, Precision, Recall, F1-score, Specificity, ROC-AUC and Confusion Matrix. This diagram reflects a systematic pipeline combining both unsupervised learning (GMM) with SMOTE along with Optimized XGBoost to enhance BD detection from complex clinical data.

3.3. Addressing Class Imbalance

Given the imbalance between bipolar and non-bipolar cases, SMOTE was employed to synthetically generate minority class instances within the training data, ensuring a balanced representation of both classes.

3.4. GMM Clustering and Hybrid Feature Engineering

A Gaussian Mixture Model with four components was trained on the balanced feature space to uncover latent patient subgroups. Each subject was assigned a cluster label which was then concatenated as an additional feature to the existing clinical dataset. Figure 2 illustrates the sequential workflow of the proposed architecture of system.

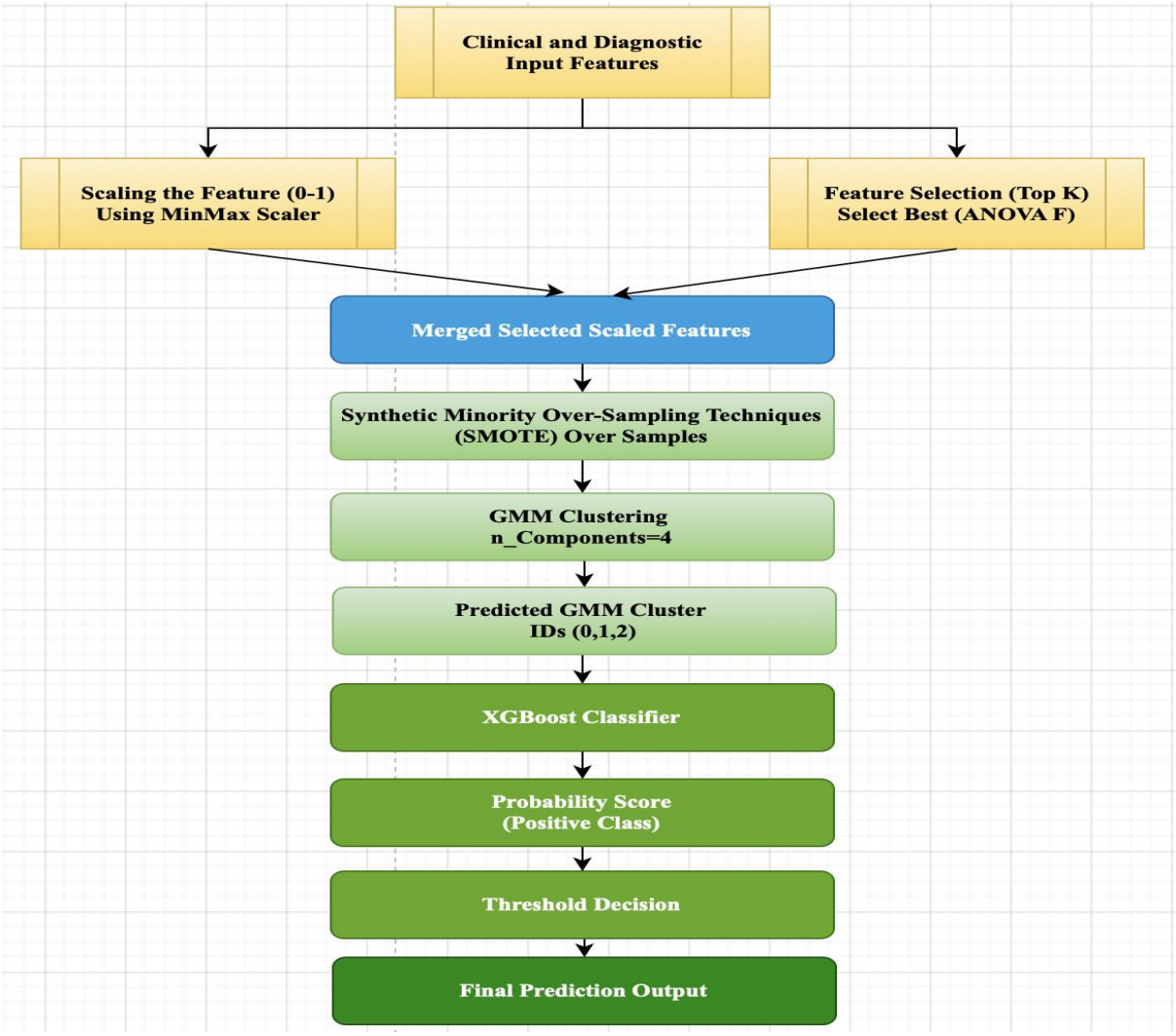

Figure 2. System Architecture Overview

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.5. Model Training and Evaluation

The enhanced feature set was used to train a GMM+SMOTE+Optimized XGBoost classifier with balanced class weights. The model's performance was evaluated using Accuracy, Precision, Recall, F1 Score, ROC-AUC, and Specificity. The proposed approach was compared against LT, DT and SVM to validate the model's effectiveness. Figure 3 shows the workflow of the Layered Architecture of the (GMM+SMOTE+Optimized XGBoost) Model.

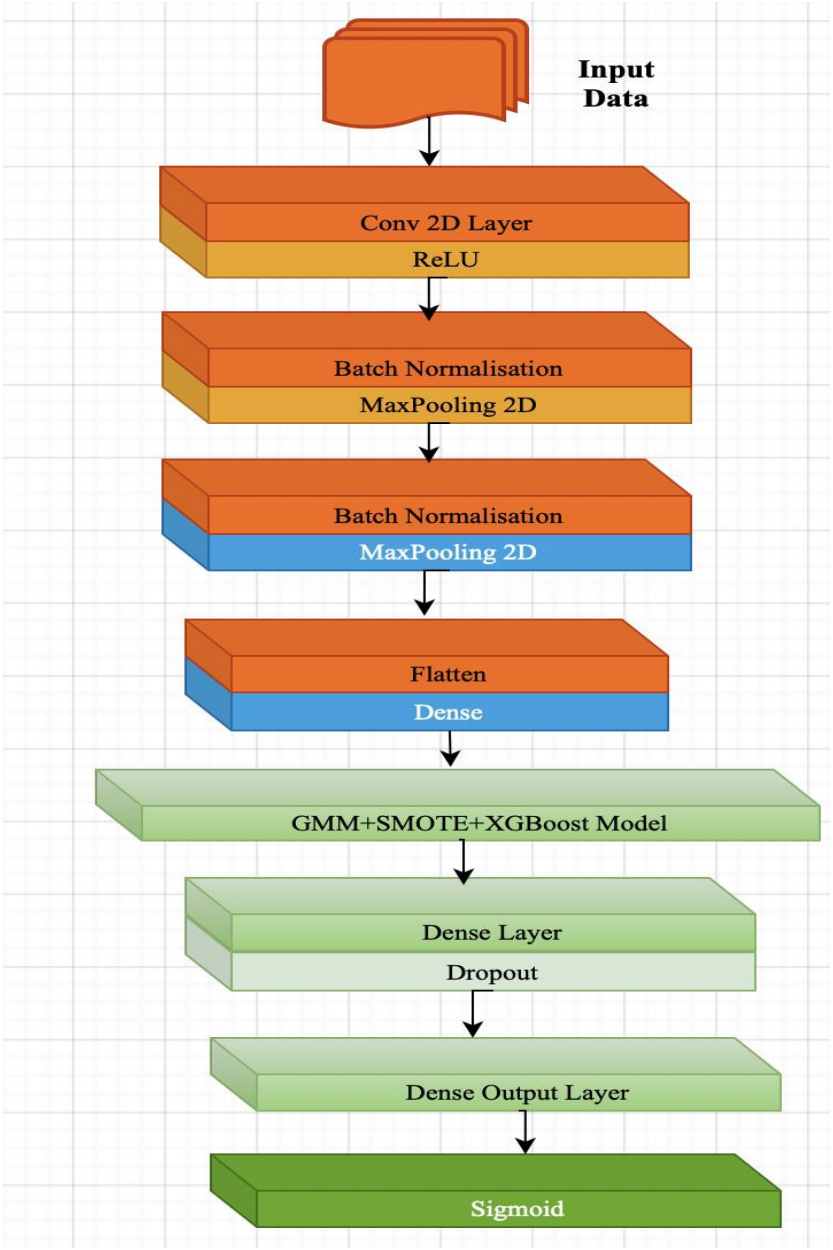

Figure 3. Layered Architecture of the (GMM+SMOTE+Optimized XGBoost) Model

A. Pseudocode for the model

BEGIN

1. LOAD DATASET

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Read the dataset (features + target)

- Split into X (features) and y (labels)

2. TRAIN-TEST SPLIT

- Stratify split to preserve class distribution

3. FEATURE SCALING

- Apply Min-Max Scaling to bring features in [0,1] range

4. UNSUPERVISED CLUSTERING (GMM)

- Fit a Gaussian Mixture Model (GMM) with 2 components on training data

- Predict cluster membership probabilities (soft clustering)

- Append these cluster probabilities as additional features to X_train and X_test

5. CLASS BALANCING (SMOTE)

- Apply Synthetic Minority Oversampling Technique (SMOTE)

→ Oversample the minority class in the training set

- Output balanced training data (X_train_balanced, y_train_balanced)

6. DEFINE BASE XGBOOST MODEL

- Set basic parameters: binary:logistic objective, random_state, etc.

7. HYPERPARAMETER OPTIMIZATION

- Define a parameter grid:

- n_estimators ∈{100, 200, 300}

- learning_rate ∈{0.01, 0.05, 0.1}

- max_depth ∈{3, 5, 7}

- subsample ∈{0.8, 1.0}

- colsample_bytree ∈{0.8, 1.0}

- Perform GridSearchCV with 5-fold cross-validation

- Select the best hyperparameters (best_xgb)

8. TRAIN FINAL XGBOOST MODEL

- Train best_xgb on balanced training data

9. TESTING & PREDICTION

- Predict class labels on X_test_combined

- Predict probabilities for ROC-AUC

10. EVALUATION

- Compute:

- Accuracy

- Precision

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Recall

- Specificity

- F1-score

- ROC-AUC

- Print results

11. PLOT ROC CURVE

- Plot False Positive Rate vs True Positive Rate

- Show AUC

END

B. Algorithm for the model

Step1. Data Preprocessing:

1.1: Handle missing values in X (e.g., mean or median imputation)
1.2: Normalize features using Min-Max Scaling to transform all values into the range [0,1]

1.3: Apply Feature Selection (SelectKBest with ANOVA F-test) to retain the top K most informative features → X_selected

Step 2. Split Dataset:

2.1. Split D into Training set (X_train, y_train) and Test set (X_test, y_test) using stratified random sampling.

Step 3. Handle Class Imbalance:

3.1. Apply SMOTE to X_train and y_train to create synthetic minority class samples → balanced X_train_resampled, y_train_resampled

Step 4. Latent Cluster Extraction with GMM:

4.1. Fit a Gaussian Mixture Model (GMM) with n_components = N (e.g., 4) on X_train_resampled
4.2. Predict cluster assignments:

- *cluster_train = GMM.predict(X_train_resampled)*
- *cluster_test = GMM.predict(X_test)*

Step 5. Feature Augmentation:

5.1. Concatenate cluster_train as an additional feature to X_train_resampled → X_train_final

5.2. Concatenate cluster_test as an additional feature to X_test → X_test_final

Step 6. Model Training:

6.1. Initialize XGBoost with parameters:

- *class_weight = 'balanced'*

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- *solver = 'liblinear'*
- *max_iter = 1000*

6.2. Train XGBoost on X_train_final, y_train_resampled

Step 7. Prediction:

7.1. Predict class probabilities on X_test_final → y_prob

7.2. Apply decision threshold (e.g., 0.4) to generate final predictions → y_pred

Step 8. Performance Evaluation:

8.1. Compute the following metrics using y_test and y_pred:
- *Accuracy*
- *Precision*
- *Recall (Sensitivity)*
- *Specificity*
- *F1 Score*
- *ROC-AUC Score*

Step 9. Comparison with Other Models:

9.1. Train and evaluate baseline models LR, DT, and SVM using the same training and testing sets.
9.2. Compute the same metrics for all models.

Step 10. Visualization & Reporting:

10.1. Plot comparative bar charts for all models across all metrics.
10.2. Report the best-performing model based on overall accuracy, F1 Score, and ROC-AUC.

3. Results & Discussion

3.1. Performance Metrics for the Model

The GMM-SMOTE-XGBoost hybrid model achieved superior classification results on the independent test set. To assess the effectiveness of the GMM-SMOTE-XGBoost hybrid model for BD detection, a comprehensive performance evaluation was conducted using multiple classification metrics. The evaluation focused on both the predictive accuracy and the model's ability to generalize across balanced and imbalanced datasets.

3.1.1. **Accuracy (ACC):**

Measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of instances:

$$Acc = \frac{C_{pos}+C_{neg}}{C_{pos}+C_{neg}+M_{pos}+M_{neg}} \quad (5)$$

Where:

- C_{pos} = Correctly predicted positive cases
- C_{neg} = Correctly predicted negative cases
- M_{pos}= Misclassified positive cases

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- M_{neg}= Misclassified negative cases

3.1.2. **Precision (Positive Predictive Value):**

Assesses the proportion of true positive predictions among all positive predictions:

$$Precision = \frac{C_{pos}}{C_{pos}+F_{pos}} \quad (6)$$

Where:

- C_{pos} = Correctly predicted positive instances
- F_{pos}= Falsely predicted positive instances

3.1.3. **Recall (Sensitivity or True Positive Rate):**

Evaluates the model's ability to identify all relevant instances of the positive class:

$$Recall = \frac{C_{pos}}{C_{pos}+F_{neg}} \quad (7)$$

Where:

- C_{pos} = Correctly predicted positive instances
- F_{neg} = Falsely predicted negative instances

3.1.4. **Specificity (True Negative Rate):**

Measures the ability to correctly identify negative instances:

$$Specificity = \frac{C_{neg}}{C_{neg}+F_{pos}} \quad (8)$$

3.1.5. **F1-Score:**

The harmonic mean of precision and recall, providing a balance between the two:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Prescision+Recall} \quad (9)$$

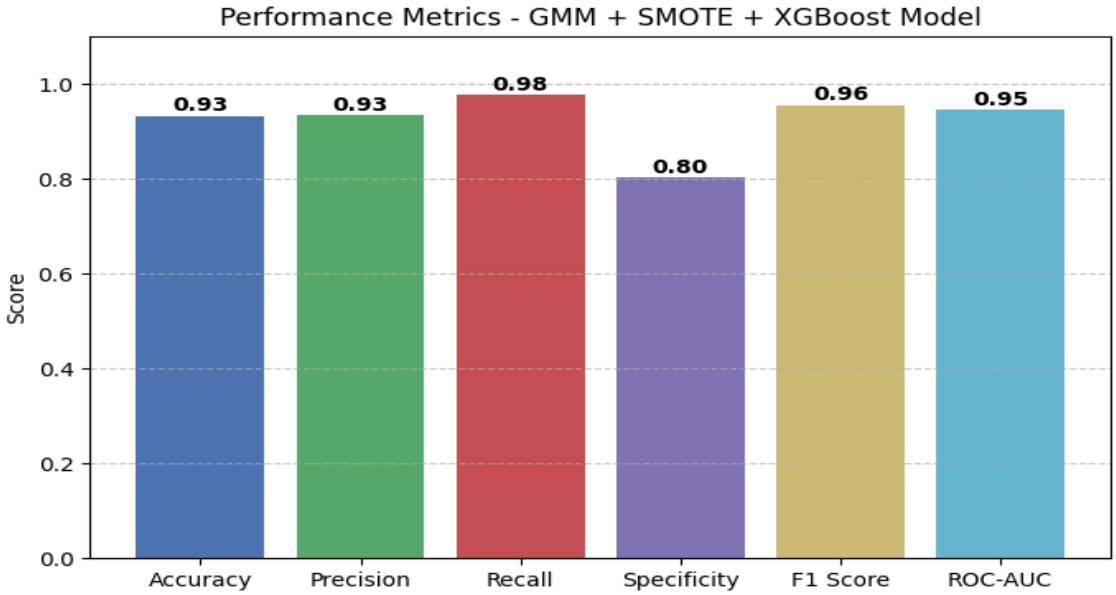

Figure 4. Performance Computation of the (GMM+SMOTE+XGBoost) Model

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2. ROC-AUC Score:

The Area Under the Receiver Operating Characteristic Curve, reflecting the model's ability to distinguish between classes across varying thresholds.

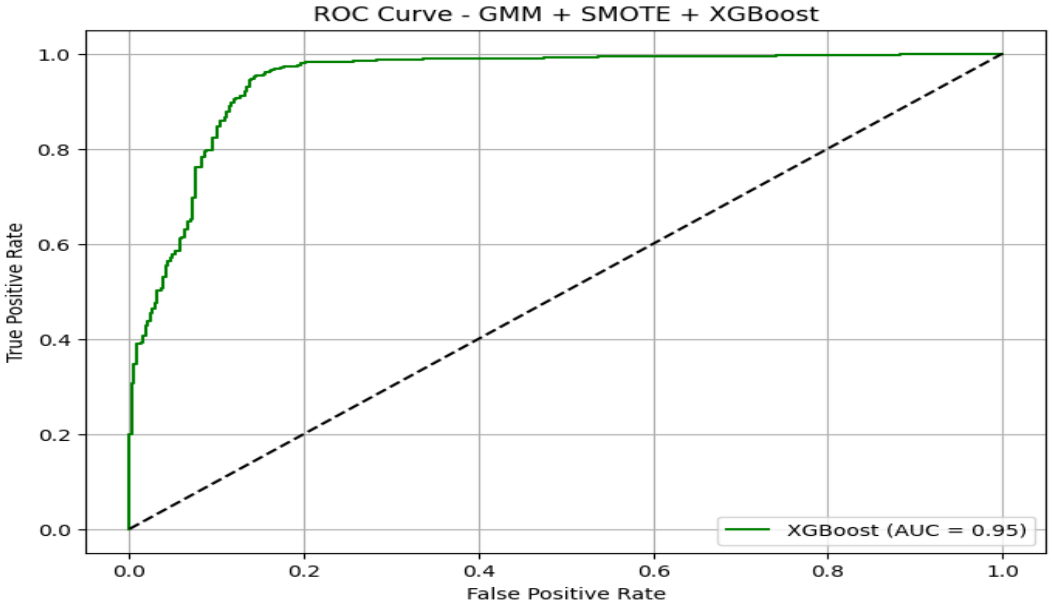

Figure 5. ROC-AUC for the (GMM+SMOTE+XGBoost) Model

3.3. Confusion Matrix:

Provides a detailed visualization of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), offering insights into the types of classification errors.

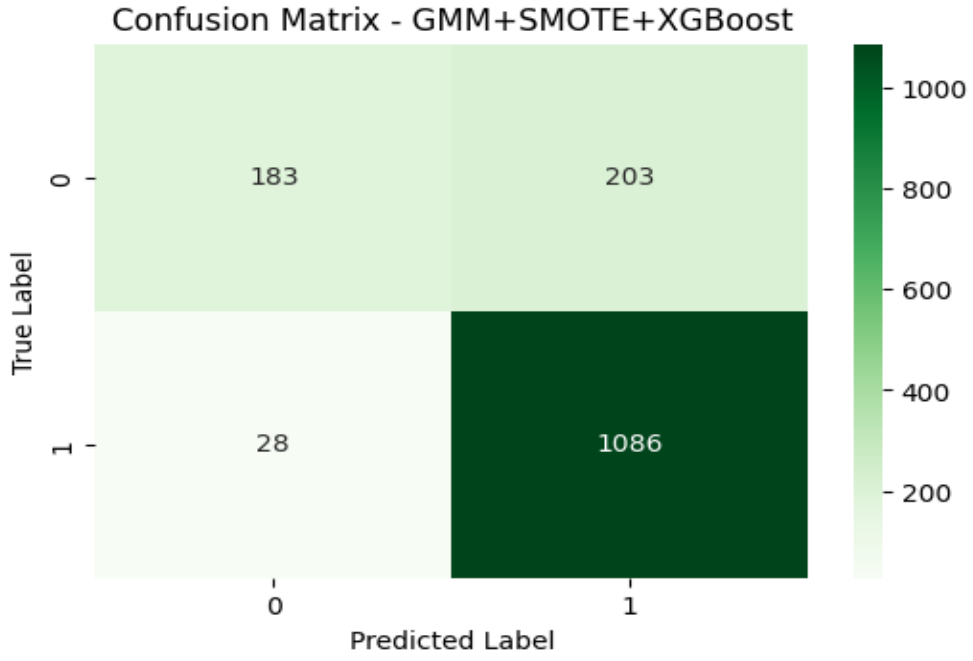

Figure 6. Confusion Matrix for (GMM+SMOTE+XGBoost)Model

The inclusion of GMM-derived cluster information provided additional discriminative power, allowing the model to better capture underlying patterns in patient data that standard classifiers

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

overlooked. Furthermore, the use of SMOTE significantly improved Recall, reducing false negatives, which is critical in mental health screening where early detection is essential.

3.4. Validation Strategy

The model was evaluated using an 80-20 train-test split with stratified sampling to ensure class distribution consistency. The following strategies were incorporated:

- Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively on the training set to correct class imbalance.
- Gaussian Mixture Model (GMM) cluster assignments were used as additional features to improve model differentiation.
- The final XGBoost classifier was trained on this enriched feature set.

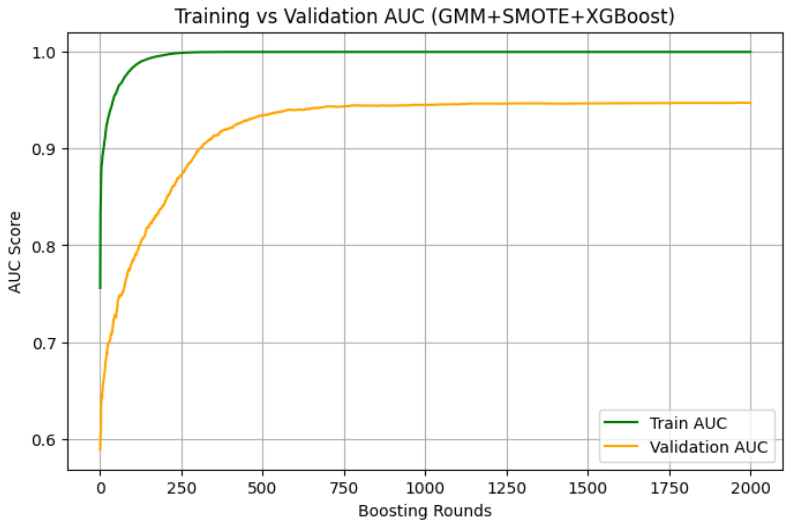

Figure 7(a). Training and Validation Curve of the (GMM+SMOTE+XGBoost) **Model**

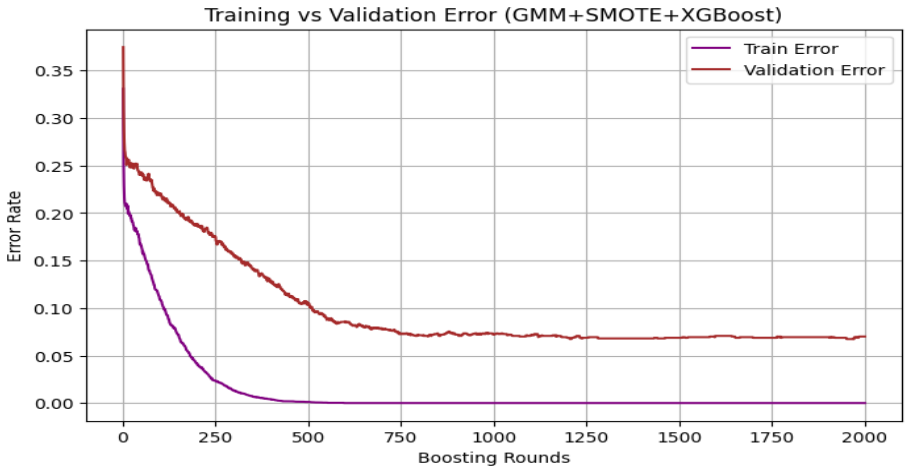

Figure 7(b). Training and Loss Curve of the (GMM+SMOTE+XGBoost) Model

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.5. Comparative Analysis

The hybrid model that is used in this research consistently outperformed all baseline models across all key performance indicators. The Accuracy of the model is 93.27% indicates strong model calibration and high separability between classes. Importantly, the F1 Score of 95.57% demonstrates a balanced trade-off between Precision and Recall, making the approach suitable for clinical decision support systems. Table 1 and figure 8 shows the performance of conventional models such as GMM+SMOTE+Optimized XGBoost, LR, DT, and SVM

Table 1: Performance Comparison Table of Multiple Models on Bipolar Dataset

Model	Accuracy	Precision	Recall	F1-Score
(GMM+SMOTE+Optimized XGBoost)	0.93	0.93	0.98	0.95
LR	0.56	0.78	0.57	0.65
DT	0.64	0.76	0.77	0.76
SVM	0.86	0.86	0.97	0.91

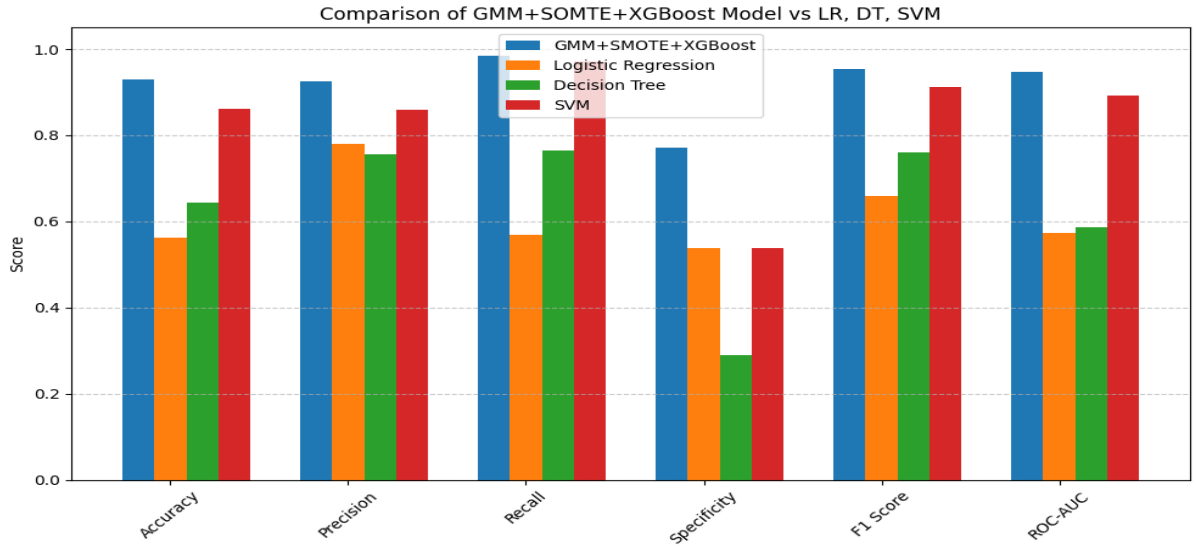

Figure 8. Comparative Analysis of the Model with other Models

4. Conclusion and Future Work

In this study, we developed a hybrid machine learning framework that integrates patient grouping using GMM with class balancing through SMOTE and final classification using an optimized XGBoost model. This GMM-SMOTE-XGBoost approach achieved high diagnostic performance for BD, with an accuracy of 93.27% and an F1 score of 95.57%, significantly outperforming conventional models such as LR, DT, and SVM. By leveraging GMM-derived cluster features, the model was able to better capture hidden patient subgroups, enhancing its sensitivity in identifying BD cases while maintaining strong specificity.

In future work, we plan to extend this framework by exploring advanced clustering techniques, such as Variational Autoencoders (VAE) or Deep Embedded Clustering (DEC), to uncover more nuanced patient subtypes. We also aim to expand the model to detect different bipolar

Received: August 05, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

subtypes and validate its robustness on larger, multi-center datasets. Finally, we envision integrating this model with real-world hospital systems and incorporating explainable AI methods to improve clinical trust and facilitate adoption in practical healthcare environments.

Conflict of interest: None authors have any conflict.

Data availability:

The datasets generated and/or analyzed during the current study are publicly available in the [https://www.kaggle.com/datasets/karanbakshi1/mental-illness-dataset].

Reference :

1. Qiu, Zhiyu, Yuxiao Hua, Tianqi Chen, Yuki Todo, Zheng Tang, Delai Qiu, and Chunping Chu. 2025. "A Gaussian Mixture Model-Based Unsupervised Dendritic Artificial Visual System for Motion Direction Detection" *Biomimetics* 10, no. 5: 332. https://doi.org/10.3390/biomimetics10050332.
2. Pal B. and Paul M. K., "A Gaussian mixture based boosted classification scheme for imbalanced and oversampled data", International Conference on Electrical, Computer and Communication Engineering, 2017, DOI:10.1109/ECACE.2017.7912938.
3. Zhang And Yang, "G-Smote: A Gmm-Based Synthetic Minority Oversampling Technique For Imbalanced Learning", Preprint, Arxiv, 24 Oct 2018.
4. Pan Z., Gui C., Zhang J., Zhu J. and Cui D. , "Detecting Manic State of BD Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech", Psychiatry Investigation 15(7), July 2018, DOI:10.30773/pi.2017.12.15.
5. Hassantabar S., Zhang J., Yin H. and Jha N. K., "MHDeep: Mental Health Disorder Detection System Based on Wearable Sensors and Artificial Neural Networks", ACM Transactions on Embedded Computing Systems, Volume 21, Issue 6, Pages 1 – 22, https://doi.org/10.1145/3527170.
6. V. A. M. Jyothi and P. S. Varma, "Exploring BD and Academic Performance with Hybrid Resnet-LSTM Model," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1082-1085, doi: 10.1109/ICUIS64676.2024.10867168.
7. P. Singh, G. Singh and S. Bharti, "The Predictive Model of Mental Illness using Decision Tree and Random Forest classification in Machine Learning," *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2022, pp. 01-05, doi: 10.1109/ICACITE53722.2022.9823761.
8. Sivagnanam L. and Visalakshi N. K., "Enhancing BD Detection Using Heterogeneous Ensemble Machine Learning Techniques", JOURNAL OF DATA SCIENCE, 2025.
9. WU C-H, HSU J-H., LIOU C-R., SU H-Y., LIN E. C-L. and CHEN P-S, "Automatic BD Assessment Using Machine Learning With Smartphone-Based Digital Phenotyping", IEEE Access, Vol. 18, 2023.
10. Pan B., Zhang H., Li D. and Yu X., "An Improved SMOTE Algorithm for Enhancing Classification Performance in Imbalanced Adolescent Mental Health Datasets", International

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Conference on Computer Applications Technology (CCAT), September 2023, DOI:10.1109/CCAT59108.2023.00037.

11. Li Y, Yang Y, Song P, Duan L, Ren R. An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. Sci Rep. 2025 Jul 2;15(1):23521. doi: 10.1038/s41598-025-09506-w. PMID: 40603552; PMCID: PMC12222711.
12. Huth F, Tozzi L, Marxen M, Riedel P, Bröckel K, Martini J, Berndt C, Sauer C, Vogelbacher C, Jansen A, Kircher T, Falkenberg I, Thomas-Odenthal F, Lambert M, Kraft V, Leicht G, Mulert C, Fallgatter AJ, Ethofer T, Rau A, Leopold K, Bechdolf A, Reif A, Matura S, Biere S, Bermpohl F, Fiebig J, Stamm T, Correll CU, Juckel G, Flasbeck V, Ritter P, Bauer M, Pfennig A, Mikolas P. Machine Learning Prediction of Estimated Risk for BDs Using Hippocampal Subfield and Amygdala Nuclei Volumes. Brain Sci. 2023 May 27;13(6):870. doi: 10.3390/brainsci13060870. PMID: 37371350; PMCID: PMC10296102.
13. Saha, D.K., Hossain, T., Safran, M. *et al.* Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks. *Sci Rep* 14, 25470 (2024). https://doi.org/10.1038/s41598-024-77193-0.
14. Kosolwattana, T., Liu, C., Hu, R. *et al.* A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Mining* 16, 15 (2023). https://doi.org/10.1186/s13040-023-00330-4.
15. Ogunseye E. O., Adenusi C. A., Nwanakwaugwu A. C., Ajagbe S. A., and Akinola S. O., "Predictive Analysis of Mental Health Conditions Using AdaBoost Algorithm", *paradigmplus*, vol. 3, no. 2, pp. 11-26, Aug. 2022.
16. Balakrishna N., Krishnan M. B. M. and Ganesh D., " Hybrid Machine Learning Approaches for Predicting and Diagnosing Major Depressive Disorder", International Journal of Advanced Computer Science and Applications, Vol. 15, No. 3, 2024.

Received: August 05, 2025