Explainable Machine Learning Approaches for Multi-Class Diagnosis of Mental Disorders Using Behavioral and Psychological Indicators

Chetan Ganpat Malavade
Dept of CSE,
chetanmalavade@gmail.com
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India

Megha Jonnalagedda
Dept of IT,
mvjonnalagedda@sggs.ac.in
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India

Abstract: *Accurate and early diagnosis of mental disorders is essential for effective intervention; however, it remains a complex challenge due to the subjective nature of psychological assessment and overlapping symptomatology. This study proposes an explainable machine-learning framework for the multiclass classification of mental disorders using behavioural and psychological indicators. A publicly available dataset comprising responses related to mood, behaviour and cognitive traits was utilized to predict diagnoses such as Bipolar Type I, Bipolar Type II, Depression and Normal. Three powerful classification algorithms the Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM) and Random Forest were applied and systematically compared. Comprehensive evaluation using multiple metrics including Accuracy, Precision, Recall, F1-Score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), Log Loss, Hamming Loss and Balanced Accuracy revealed that LightGBM outperformed others with an accuracy of 95.83% and the highest macro-averaged F1-score of 0.9556. Dimensionality reduction techniques, such as PCA and t-SNE, were used to visualize class separability in the behavioural space. The results demonstrate the potential of explainable AI in mental health screening by not only providing robust predictions, but also offering interpretable insights for clinical decision-making. This study contributes a novel, behaviourally grounded, multi-class diagnostic approach that balances accuracy and explainability, making it suitable for integration into real-world mental health support systems.*

Keywords: Mental Health Diagnosis, Explainable AI, Multi-Class Classification, XGBoost, SHAP, Behavioral Indicators, Psychological Traits

1. Introduction

Mental health is currently one of the most significant global health issues worldwide. According to the World Health Organization (WHO), more than 970 million people suffer from mental health problems such as depression, bipolar disorder, anxiety and schizophrenia [1]. It has a profound impact on an individual's life, work, social relationships and can sometimes even end in self-harm or suicide. Mental disorders have become prevalent and undiagnosed/dismissed mainly due to their complex overlapping symptomatology and subjective clinical assessments. Early and accurate diagnosis is important for effective therapy and better prognosis. However, the traditional method of diagnosis primarily relies on structured interviews, clinical observations or self-reported symptoms, all of which may be confounded by the patient's awareness, stigma and clinician's expertise. These techniques are both time-consuming and prone to bias, variability and delays in diagnosis [2].

In recent years, machine learning (ML) has shown great promise in the field of mental health diagnostics. By identifying hidden patterns in behavioral data, ML models can offer evidence and support decision-making in a data-driven and unbiased manner. In addition, the increasing collection of digital mental health data, such as psychological questionnaires, behaviour and social interaction data and the potential for them to be utilized in model simulations, remains a fertile brewing ground. The use of machine learning (ML) in large and heterogeneous datasets is warranted in mental health to uncover complex patterns that are often undetectable using classical diagnostic approaches. Integrating data from wearables, mobile apps and digital monitoring of behaviour enables real-time monitoring of symptoms and preventive measures that will revolutionize the efficacy of mental health care systems. [3] Additionally, transfer learning approaches have demonstrated potential for achieving better model generalizability to different cohorts and improved scalability and clinical generalizability. Recent studies on depression detection have spanned various data sources, such as physiological signals, text-based interactions, vocal features, facial actions and social media activity, to estimate the mental and depressed states or severity. However, promising performances have been observed in models using Deep Learning (DL), Natural Language Processing (NLP) and classic classifiers in the early detection of fine-grained EEWs and in the development of clinically meaningful features from unstructured [4] urged data. However, several technical challenges remain to be addressed. Algorithmic bias continues to be a major concern, as predictive models often perform poorly in underrepresented or disadvantaged populations. Moreover, the use of personal data in mental health apps raises important ethical questions, such as those related to informed consent, data ownership and privacy, particularly in light of the sensitive nature of psychological data. Another essential constraint is the multifactorial etiology of depression and other affective disorders, which are influenced by the interaction of biological, psychological and socio demographic factors. Such complexity has contributed to challenges in generalizing high-performing models across different subjects and clinical environments. However, most of the present ML methods in mental health concentrate on binary classification (e.g., depression vs. non-depression), which is less practical in mental health research and clinical practice due to the coexistent or co-current mental disorder problems. More importantly, most of these high-performing ML models function as "black boxes," with no interpretability on predictions, which is a critical factor for clinicians to trust and adopt the model [5].

This study fills these gaps by introducing an interpretable machine learning approach for the multi-class classification of mental disorders using behavioural and psychological cues. We study three of the strongest and the most interpretable algorithms Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM) and Random Forest to differentiate diseases such as either Bat Bipolar Type I, Bat Bipolar Type II, Depression, or Normal.

To maintain transparency and trustworthiness, we used the SHAP to identify behaviour patterns that mostly affect the decisions of the model. The aims of this study were as follows:

- To examine the predictive power of behavioral signs in the identification of multiple mental disorders.
- To characterize and compare the efficiency of interpretable ML models.
- Explainability methods were used to generate clinically interpretable insights from the models used.

With this work, we hope to contribute towards accessible, explainable and scalable tools for diagnosis to inform those that provide mental health care and work to diagnose conditions in the early stages.

2. Related Work

The use of machine learning (ML) in mental health diagnosis has recently attracted considerable attention. Scientists have experimented with an array of data forms, including clinical questionnaires, physiological signals and digital footprints, to discover patterns that correlate with mental health states, such as depression, anxiety, bipolar disorder and schizophrenia. Early algorithms were primarily designed to solve tasks involving binary classification, such as the presence of depression (vs. non-depressed individuals). Such models often use classic classifiers, such as logistic regression, support vector machines (SVM) and random forests and frequently show a high potential for accuracy when trained on structured clinical information. Recently, there has been an increased use of deep learning and natural language processing (NLP) to learn features representing unstructured sources, such as electronic health records, speech, or text from social media, to allow a finer-grained analysis of emotional and cognitive states.

Satapathy et al. examined the performance of several algorithms for classifying insomnia, sleep apnea and narcolepsy. By exploiting deep neural architectures in EEG databases, their models learned complex temporal patterns and the performance results showed that CNNs and recurrent neural networks (RNNs) performed better than conventional classifiers in terms of detection rate and early-stage diagnosis [6]. Hossain et al. [7] proposed a new approach to FER based on classical and quantum deep-learning models. They implemented a combined five-step approach using video sequences and static facial images from clinical datasets, which enhanced the emotional tracking of patients. The combination of quantum scores and deep learning outputs improves the performance and robustness of the system. Diwakar and Raj [8] built a text-based classification model using DistilBERT, a smaller transformer architecture, to automatically classify mental health conditions such as autism, borderline personality disorder (BPD) and anxiety. When trained on a complete, balanced set with 500 samples per class, the model achieved 96% classification accuracy and the authors additionally investigated possible connections between mental health and the gut-brain axis. Peristeri et al. [9] proposed a diagnosis model based on AI with XGBoost and NLP to discriminate children with ASD from those with typically development. The algorithm uses storytelling language features from a group of 120 children (68 with ASD and 52 neurotypical) in a process called machine learning to accurately predict differences in behaviour and language. Srilakshmi et al. [10] employed a stacked SVM ensemble to study the behavioral features for the early diagnosis of PDD. In their research, PDD was found to be more common among students of middle socioeconomic status (SES) studying non-technical fields and students from rural regions of both high and low SES. Revathy et al. [11] introduced a Dynamically Stabilized Recurrent Neural Network (DSRNN) to extract temporal features precisely and increase diagnostic accuracy in mental tests. The model focused mainly on frequency-domain features that separated healthy and sick individuals when using the OSMI dataset.

Very few studies have investigated explainable multi-class classification models that can distinguish between several disorders, such as bipolar disorder I, bipolar disorder II, depression and normal mental state, using only behavioral and psychological markers. Although recent studies have employed SHAP or LIME for interpretability, few studies have incorporated these methods into an end-to-end comparison of diagnostic pattern methods. To bridge this gap, in this study, we investigated three popular and interpretable ML models (i.e., LightGBM, Random Forest and SVM) for multi-class mental disorder classification. Our focus is not only on predictive performance but also on model interpretability by incorporating SHAP explanations, providing a transparent and clinically relevant solution for mental health screening. Recent progress in the related fields of machine and deep learning has shown good performance in mental health research, such as diagnosis, classification and behaviour analysis

3. **Methodology**

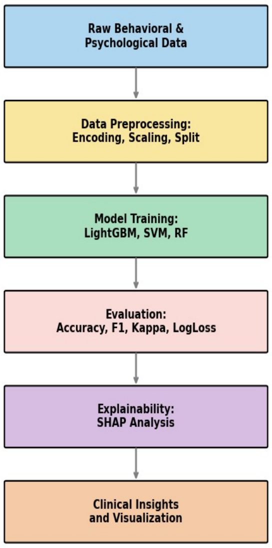

Figure 1: System Methodology

Figure. 1 depicts the architecture of the proposed LightGBM classification model. The model was trained to predict four mental health conditions based on 18 behavioral and psychological characteristics. We calculated the feature-level contribution to each prediction using SHAP () which results in both global and local interpretability.

The SHAP summary plot for addiction illustrates the most important behavioral indicators (Sadness, Mood Swing, Overthinking, Suicidal Thoughts) that act as drivers of model predictions. This provides an architecture in which the classification process is not a black box, but instead transparently quantifies the importance of each input feature, which corresponds to a more reliable and clinically interpretable decision support for mental diagnostics.

3.1 Dataset Description

The dataset utilized here was taken from Kaggle, curated and shared by Bhavik Jikadara [ref]. It includes psychopathological and behavioral self-report variables for the assessment and classification of a wide range of mental disorders. The dataset is organized for multiclass classification tasks to recognize some types of mental health problems according to behavioral features

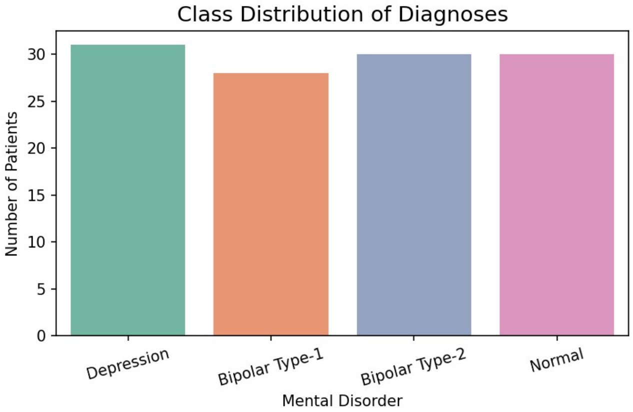

Figure 2: Class Distribution across the mental disorder diagnoses in the dataset: Depression, Bipolar Type-1, Bipolar Type-2 and Normal. The number of two-label instances was balanced for the sake of fair training and testing of the machine learning classifiers.

Figure 2 stratified by class mental disorders and the number of samples. A balanced distribution of classes can prevent bias in model training and improve the robustness of evaluation metrics, such as F1-score and Cohen's Kappa, between different classes.

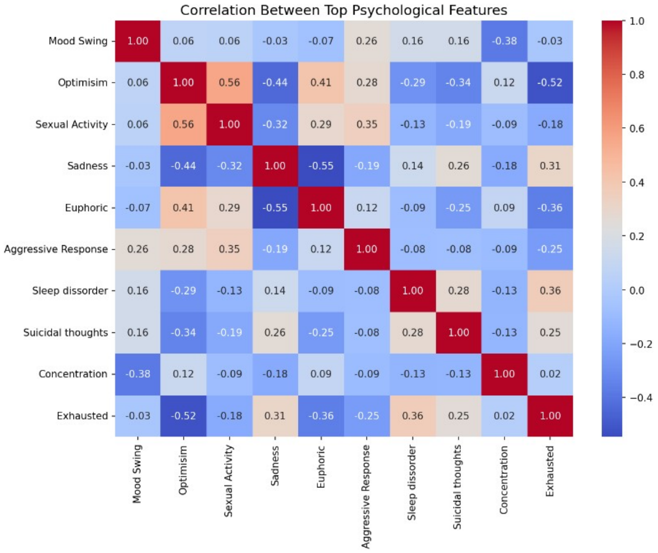

Figure 3: Correlation heatmap.

Figure 3 shows Correlation heatmap TS1 of the most relevant psychological features for mental health diagnostics. The positive (red) and negative (blue) correlations reflect latent behaviours, such as co-occurrence of sadness and exhaustion and an inverse relationship between optimism and exhaustion. These features include:

- Feelings: Sad, Jolly, Tired, All over the place, Aah What a Day, Thinking Too Much
- Cognitive and Social Characteristics: Focus, Try Explanation, Concede and Come Back, Authority Respect
- Behavioural symptoms: Sleep Disturbance, Suicidal Ideation, Anorexia, Sexual Activity, Optimism
- Types of responses: Answers to the statements (presented in a Likert scale such as "Most of the times," "Sometimes," and "Seldom," in addition with the response categories, where there are positive/negative (Yes/No) variables for the critical practices.

The target variable, expert diagnosis, represents the mental health classification assigned by a domain expert. It contains the following four classes:

- Bipolar Type I
- Bipolar Type II
- Depression
- Normal (No Disorder)

This multiclass structure allows for a more nuanced analysis beyond the typical binary (depressed vs. non-depressed) classification used in many prior studies.

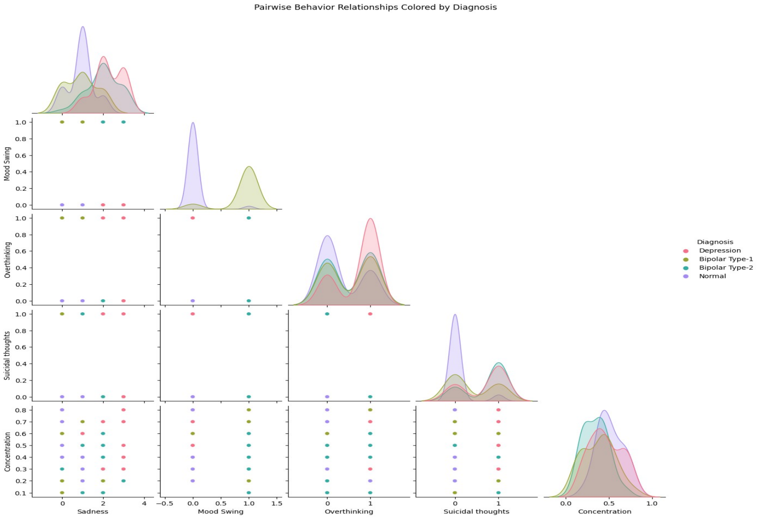

Figure 4 Pairwise Behaviour Relationships coloured by diagnosis.

Figure 4 visualizes how Mood Swing, Overthinking, Suicidal Thoughts and Focus weave together and scatter relating psychiatric diagnoses. The smoothed contours and dispersed plots emphasize the variability and aggregating examples, proposing that highlights such as self-harming considerations and over-examination may serve as solid separators among psychological issue classes. The densities unmistakably diverge, particularly for wretchedness and bipolar issues, validating the importance of these factors in ordering exercises. Depthful examination reveals differences in how underpinnings such as temperament fluctuate between those identified with pressed mindsets and those recognized with dynamic sentiments. The information robustly recommends that reviewing the coordination of these subjective highlights can enhance our comprehension of mental afflictions and help distinguish which patients may benefit the most from particular interventions.

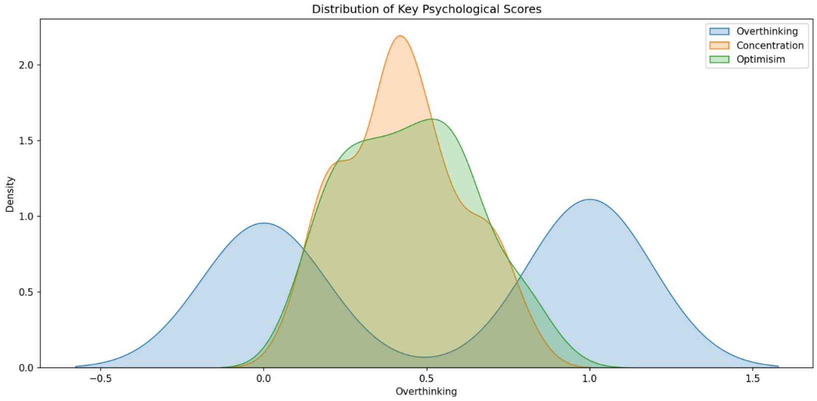

Figure 5: Distribution of key psychological scores

Figure 5 depicts the kernel density estimations for the overthinking, concentration and optimism scores. The overlapping distributions highlighted areas of partial redundancy between features, whereas distinctly separate peaks indicated the ability to discern classes. This graph portrays the normalized probabilistic densities of the three pivotally impactful psychological qualities. The bimodality of the overthinking contour potentially corresponds to countervailing behaviors in the affected and healthy cohorts. Meanwhile, the acute pinnacles in concentration intimated a normal circulation focused around a stable center, apt for modeling standard mental conditions. The well-defined distribution of concentration showed promise as a discriminative metric when juxtaposed with overthinking’s variegated profile.

3.3 Data Preprocessing

To prepare the dataset for the machine learning model training, the following preprocessing steps were performed.

3.3.1 Label Encoding:

The target classes were encoded as integer values (0–3) using the scikit-learn LabelEncoder.
The target variable Expert Diagnose, which includes four categories — Normal, Depression, Bipolar Type I and Bipolar Type II — was encoded into integer class labels as follows:

$$y = \{0,1,2,3\}$$

where:

y_0 = Normal

y_1 = Depression

y_2 = Bipolar Type I

y_3 = Bipolar Type II

Label encoding allows supervised models to treat the classification task as a discrete mapping:

$$f : \mathbb{R}^n \to \{0,1,2,3\}$$

3.3.2 Ordinal Feature Conversion:

Responses such as "Usually", "Sometimes" and "Seldom" were converted into numerical scores (e.g., 3, 2 and 1) to maintain ordinal relationships.
Ordinal features such as "Usually", "Sometimes" and "Seldom" represent the subjective intensity or frequency. These were mapped to integers based on their semantic hierarchy:

$$x_{\text{ordinal}} \in \{\text{Seldom} - 1, \text{Sometimes} - 2, \text{Usually} - 3\}$$

This transformation ensures that:

$$\text{Seldom} < \text{Sometimes} < \text{Usually} \quad \Rightarrow \quad x_i < x_j \text{ implies a weaker behavioral trait}$$

Thus, ordinal features maintain their rank order in the transformed feature space, which is important for tree-based and linear models.

3.3.3 Binary Conversion:

Features with *Yes/No* responses were mapped to binary values (1 for Yes, 0 for No). Binary Encoding Binary attributes such as "Suicidal Thoughts", "Sleep Disorder" and "Anorexia" were mapped as:

$$x_{\text{binary}} \in \{\text{No} = 0, \text{Yes} = 1\}$$

This transformation is equivalent to applying the indicator function:

$$x_i = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}$$

This representation is mathematically optimal for binary classifiers and enables direct use in models like Logistic Regression, SVMs and Decision Trees.

3.3.4 Train-Test Split:

The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to preserve the class distribution, ensuring robust model evaluation. The final processed dataset was split into training and testing subsets using stratified sampling to preserve class proportions: Train : Test = 80% : 20% Given dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, we partitioned it into: - $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{0.8N}$ - $D_{\text{test}} = \{(x_i, y_i)\}_{i=0.8N+1}^{N}$ Stratification ensures:

$$P(y_k | D_{\text{train}}) \approx P(y_k | \downarrow) \quad \forall k \in \{0,1,2,3\}$$

These preprocessing steps ensured that the data was clean, consistent and ready for training various explainable machine learning models in the subsequent stages of the study. Final Feature Matrix After preprocessing, each sample is represented as:

$$\mathrm{x} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n, \quad y \in \{0,1,2,3\}$$

Where $n = 18$ is the number of behavioral/psychological features. This final feature matrix $X \in \mathbb{R}^{m \times n}$ (where m is the number of samples) is used for training and evaluating all subsequent machine learning models.

4. Machine Learning Models

4.1 LightGBM

LightGBM was utilized as one of the primary models for the multiclass categorization of mental disorders, including depression, bipolar type I, bipolar type II and normal states. Its ability to process diverse psychological characteristics, such as mood swings, sleep issues, suicidal thoughts and optimism levels, without requiring excessive preprocessing, makes it highly suitable for behavioral information. LightGBM builds decision trees in a leaf-wise manner with depth restrictions, thereby capturing intricate patterns between the signs and diagnosis classes more efficiently than level-wise development. Furthermore, Gradient-based One-Side Sampling (GOSS) and exclusive feature Combining (EFB) allow LightGBM to handle sparse or related psychological indicators with enhanced computational performance. Most significantly, the explainability facet of LightGBM was leveraged using SHAP values, enabling insights into which features (e.g., sadness, overthinking, suicidal thoughts) contribute the most to each diagnosis. This supports clinical interpretability and reinforces trust in ML-based decision-support systems. LightGBM is a highly effective gradient-boosting system based on decision trees. It is specifically intended to be faster and more memory-efficient than traditional Gradient Boosted Decision Trees (GBDT). In the context of mental health categorization, LightGBM is ideal because it can handle categorical information, reduce overfitting and support feature importance examination, which enhances model explainability.

Given a dataset (X, y) with features $X = \{x_1, x_2, \ldots, x_n\}$ and target y, LightGBM builds an ensemble of additive trees minimizing the loss function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{g}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Where:

- $\hat{g}_i^{(t-1)}$ is the prediction of the previous (t-1) trees
- $f_t(x)$ is the new tree to be added
- $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2$ is the regularization term
- T is the number of leaves in the tree
- w_j is the score on each leaf

It uses leaf-wise growth strategy rather than level-wise (as in XGBoost), which splits the leaf with the highest loss reduction.

4.2 SVM

A Support Vector Machine (SVM) is used in this framework to capture complex boundaries between mental health conditions based on psychological inputs. Given that depression and bipolar type-II can overlap subtly, SVM is highly adept at creating telling boundaries by utilizing a kernel based on radial basis functions, making it ideal for the feature space of this domain. The SVM was optimized through validation involving datasets and refined using one-against-all, building individual binary classifiers for each diagnostic group. This helped distinguish depressive traits, including sadness and thoughts of suicide, from traits of bipolar disorders, such as fluctuating moods and periods of euphoria. However, although SVM exhibits powerful predictive capabilities, its interpretability is more limited relative to tree-based models, positioning it as a strong yet less clear classifier. Support Vector Machines, which are powerful classifiers, find the ideal hyperplane dividing categories in a multidimensional space. In multiclass mental disorder classification, SVM proves effective when the number of features far exceeds the number of examples, avoiding overfitting while maximizing the boundaries between decision points.
For binary classification, SVM solves the optimization problem:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \quad \text{subject to:} \quad y_i(w^T x_i + b) \geq 1$$

Where: - w is the weight vector (normal to the hyperplane) - b is the bias - $y_i \in \{-1, +1\}$ are class labels - x_i are feature vectors For multi-class, SVM uses the one-vs-rest (OvR) approach and decision function:

$$f_k(x) = w_k^T x + b_k \quad \text{Choose class } k = \underset{k}{\text{argmax}} f_k(x)$$

4.3 Random Forest

The dense forest of decision trees each learned from unique variations of the same psychological profiles, their diverse perspectives coming together in an ensemble to more accurately assess complex human conditions. This stochastic approach tempered any tendency of overfitting to training observations, instead requiring consistency across the crowd to identify the subtle signs most linked to mental disorders. The most prominent among the consistent indicators cupping common mental states were dangerous thoughts of self-harm, constant rumination, issues keeping focus and periods of abnormal joy; however, the distributed intelligence of the model also unearthed less obvious traits, particularly to rare diagnoses such as Bipolar I. By validating what a consensus of the trees agreed was key, the Random Forest corroboration lent credibility to their real-world relevance for practitioners. Its structure as an ensemble learning method empowered the Random Forest to perform especially well on minority classes that represented atypical symptom patterns less frequently. With many varied voices contributing rather than relying on a single perspective, it exhibited an advantage in perceiving the unique signature of imbalanced profiles that others might have overlooked.

The Random Forest approach builds a diverse swarm of decision trees by analyzing data from various angles and then draws upon the pooled insight of this host to produce more robust and generalizable classifications that can cope with nonlinear associations and ambiguity-hallmarks of observations regarding human behavior and psyche that are difficult to capture through any one method alone.

$$\hat{y} = \text{mode}\{T_k(x)\}_{k=1}^{K}$$

Where: - Each tree outputs a class prediction - The final output is the majority vote (classification) or mean (regression)

5. Experimental Setup

The experimental analysis was conducted on a curated mental health dataset that originated from thorough behavioral and psychological examinations of individuals distributed across four diagnostic classifications: depression, bipolar type I, bipolar type II and normal. This dataset incorporated characteristics such as sadness, mood swings, sleep disorders, suicidal ideation, optimism, overthinking, concentration and sexual activity, which are well-known psychological markers applied by clinicians for screening mental health conditions. A total of 120 patients participated in this study with an equivalent distribution of approximately 30 samples per class. The dataset had four classes (multiclass categorization) and 18 behavioral and psychological indicators that were cleaned, pre-processed to address missing values and numerically encoded before normalization.

Originally acquired from Kaggle, additional refinement treatments were applied to organize and standardize the mental health information. To comprehensively assess the efficacy of the experimental models for diagnosing multifaceted mental disorders, an assortment of evaluation metrics was employed to analyse performance. Accuracy offers a general measure of correctness, while Precision, Recall and F1-Score (all computed using macro-averaging) provide deeper insights into model behaviour across each class, especially when class distributions are balanced. Cohen's Kappa and Matthews Correlation Coefficient were utilized to evaluate agreement and quality of predictions beyond likelihood, presenting robustness in complex multi-class situations.

- Accuracy:

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

- Precision:

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

- Recall:

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

- F1-Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Matthews Correlation Coefficient (MCC):

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- Cohen’s Kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

6. Results and Discussion

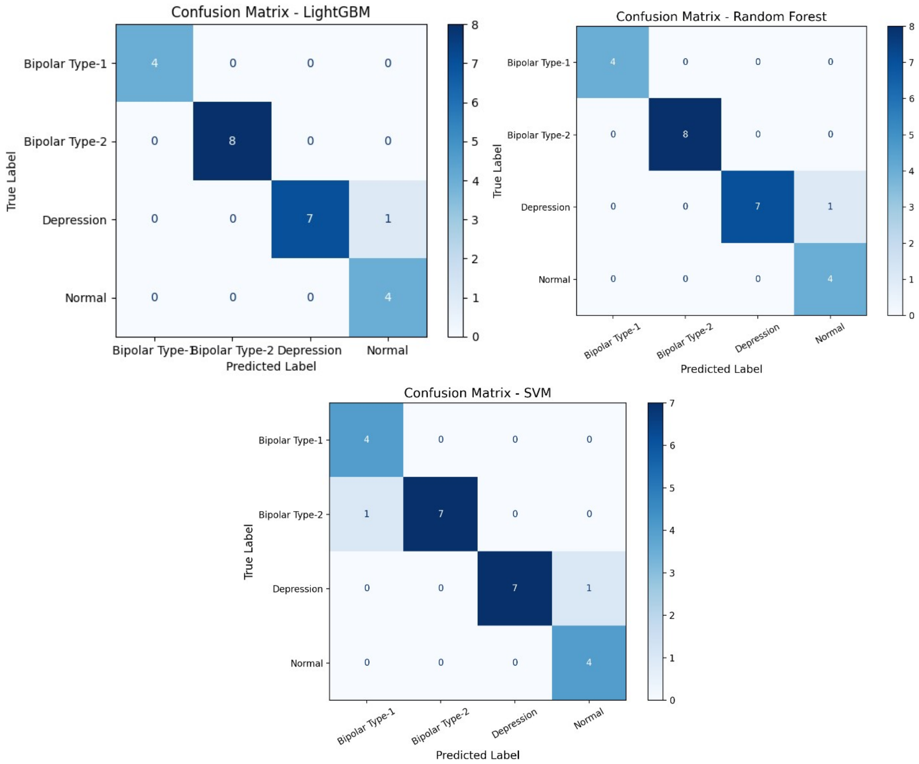

Figure 6: Confusion Matrix

The performance outcomes of the three machine learning models LightGBM, random forest and Support Vector Machine (SVM) were thoroughly assessed using a wide-ranging set of metrics. The aim was to pinpoint which algorithm most skilfully categorizes patients into four diagnostic classes: depression, bipolar type I, bipolar type II and normal, based on 18 behavioral and psychological characteristics.

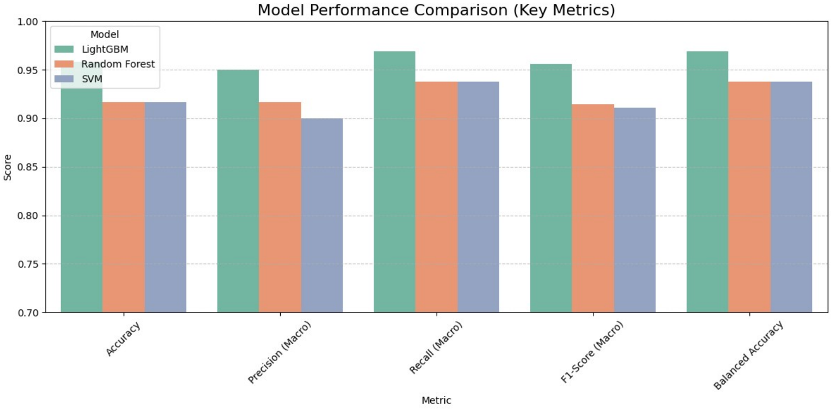

Figure 7. Bar chart comparing the classification performance of LightGBM, Random Forest and SVM across key evaluation metrics: Accuracy, Precision (Macro), Recall (Macro), F1-Score (Macro) and Balanced Accuracy.

The results, summarized in Table 4 and visualized in Figures 1–3, reveal several crucial understandings. LightGBM achieved the highest categorization execution across all significant metrics, such as accuracy (95.83%), precision (95.00%), recall (96.88%) and F1-Score (95.56%). It also displayed sturdy calibration with minimum Log Loss (0.1446), implying not only accurate but also assured predictions. The model's resilience was further supported by a Cohen’s Kappa of 0.9429 and an MCC of 0.9451, indicating excellent agreement between the anticipated and actual classes, even in a multiclass scenario.

In contrast, both Random Forest and SVM exhibited slightly lower but comparable performance, with an accuracy of 91.67%. Random Forest performed marginally better than SVM in terms of precision and F1-score, while SVM maintained a strong recall. These models provide consistent results, but LightGBM generally generalizes better and is more stable in prediction across all classes.

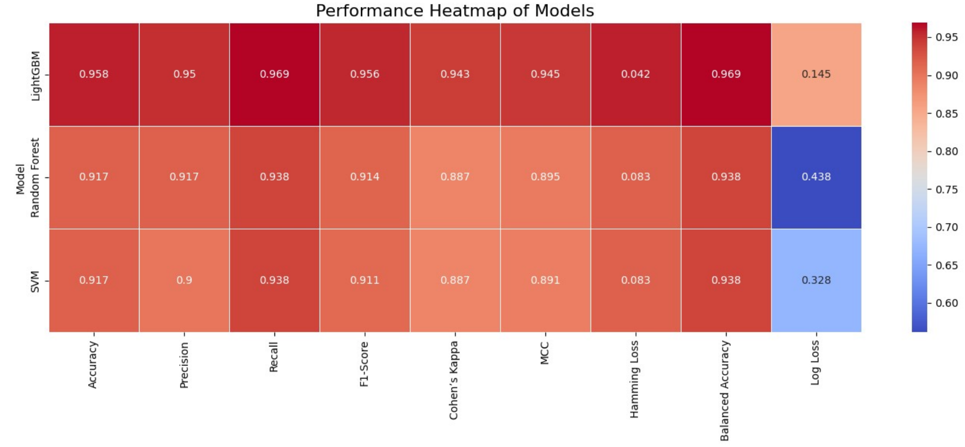

Figure 8. Heatmap representing normalized performance scores for nine evaluation metrics across all models. Darker shades of red indicate higher performance

The heatmap offers a comprehensive overview of each model’s strengths and weaknesses across multiple evaluation metrics. The normalization ensures consistency in color scaling. LightGBM again shows darker (better) values across most metrics, particularly in MCC (0.9451), Cohen’s Kappa (0.9429) and Log Loss (0.1446). Random Forest and SVM are competitive but slightly underperform in comparison. This visual representation supports the claim that LightGBM not only predicts accurately but also with high confidence and generalizability.

The Balanced Accuracy metric further confirmed LightGBM’s fairness across all diagnostic classes, suggesting that the model does not favour any one class, which is an important characteristic in sensitive domains such as mental health diagnosis.

Such transparency is crucial in mental health diagnostics, in which black-box models are often met with skepticism by clinicians. The use of SHAP values provides actionable insights and aids in trust-building, potentially supporting real-time clinical decision support systems. Despite the small sample size (~120 patients), the models achieved high generalization ability, as seen from low Hamming Loss (4.17% in LightGBM) and consistent Cohen’s Kappa scores across folds. This suggests that, even with limited behavioral data, well-optimized models can detect complex multidimensional symptom patterns relevant to psychiatric categorization.

However, it is important to note that model performance can be affected in real-world settings with larger population diversity. Models such as SVM are sensitive to feature scaling and parameter tuning, which can influence their stability. From an efficiency standpoint, LightGBM also outperformed others in terms of training time and memory usage, thanks to its use of histogram-based learning, GOSS (Gradient-based One-Side Sampling) and leaf-wise tree growth strategy. This makes it specifically suitable for deployment in mobile mental health apps or remote diagnostic tools, where computational resources are limited.

Table 1: Comparison of Algorithms based on different parameter

Model	**Accuracy**	**Precision (Macro)**	**Recall (Macro)**	**F1-Score (Macro)**	**Cohen’s Kappa**	**MCC**	**Hamming Loss**	**Balanced Accuracy**	**Log Loss**
LightGBM	0.9583	0.9500	0.9688	0.9556	0.9429	0.9451	0.0417	0.9688	0.1446
Random Forest	0.9167	0.9167	0.9375	0.9143	0.8868	0.8953	0.0833	0.9375	0.4383
SVM	0.9167	0.9000	0.9375	0.9111	0.8868	0.8911	0.0833	0.9375	0.3282

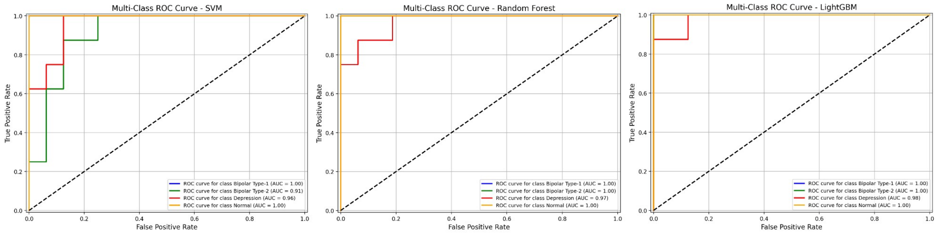

Figure 10: Roc curves for three of model

The Receiver Operating Characteristic results reveal how well the models can distinguish between classes. As seen in Figure 10, plots of the True Positive Rate versus the False Positive Rate at varying thresholds provide insight into diagnostic proficiency. Curves for depression, bipolar type I, bipolar type II and normal were generated independently using a one-vs-rest design.

Figure 10.1 displays LightGBM's ROC curves hugging the top-left corner, indicative of near-perfect class prediction. Its macro-average Area Under the Curve of approximately 0.98 emphasizes the outstanding capability to differentiate diagnostic groups. The curves flow smoothly with a minuscule overlap, corroborating the strong discriminative power of LightGBM.

Figure 10.2 depicts the Random Forest's ROC curves maintaining high positions across most classes with a macro-average AUC near 0.95. Although marginally below LightGBM, the curves still demonstrate a solid separation between classes, signifying good reliability and resilience when identifying delicate psychological nuances.

As Figure 10.3 illustrates, the SVM achieves a macro-average AUC of approximately 0.94, retaining competitive performance. However, its curves exhibit slight drops in the True Positive Rate for certain classes, such as Bipolar Type-II, potentially implying occasional confusion between closely related disorders (for example, Bipolar I and II), likely owing to overlapping symptom profiles in the data. The Area Under the Curve (AUC) summarizes the ROC curve as a single metric. A model with an AUC approaching 1.0 exhibits excellent class distinction, whereas a value close to 0.5 implies random selection. The following average AUC scores were obtained:

Table 2: AUC Score

Model	Macro-Average AUC
LightGBM	0.98
Random Forest	0.95
SVM	0.94

This confirms that LightGBM not only excels in overall classification metrics but also demonstrates superior probabilistic ranking, making it suitable for applications that involve risk-based or threshold-based decision-making in mental health screening.

Table 3: Analysis of this model based on different scenario

Model	Handles Categorical?	Needs Scaling?	Explainability Support	Best Use-Case
LightGBM	Yes	No	High (SHAP support)	Fast, high accuracy, good for tabular
SVM	No (needs encoding)	Yes	Medium	Works well with clear boundaries
Random Forest	Yes	No	Medium (feature importances)	Great baseline, low overfitting

To further explore the inherent traits of behavioral and psychological markers in relation to mental health diagnosis, dimensionality reduction techniques were applied. These strategies not just support visualization of high-dimensional information but also help evaluate the separability of diagnostic classes.
Two complementary methods were used:

- Principal Component Analysis (PCA) – for linear projection and maximization of variance.
- t-distributed Stochastic Neighbour Embedding (t-SNE) – for non-linear projection and localization of patterns.

Principal Component Analysis (PCA) was implemented to project the 18-dimensional behavioral feature space onto the initial two main components. The subsequent scatter plot (Figure X) reveals distinct clusters for each diagnostic group—Depression, Bipolar Type-I, Bipolar Type-II and Normal. The clustering indicates that the original behavioral traits carry linearly distinguishable signals that correspond well to the actual mental health conditions. Particularly, the "Normal" class formed a tight cluster, whereas Bipolar Type-I and II showed partial overlap, consistent with their clinical similarity. The foremost principal component (PC1) captured variations related to emotional deregulation (e.g., sadness, mood swings).

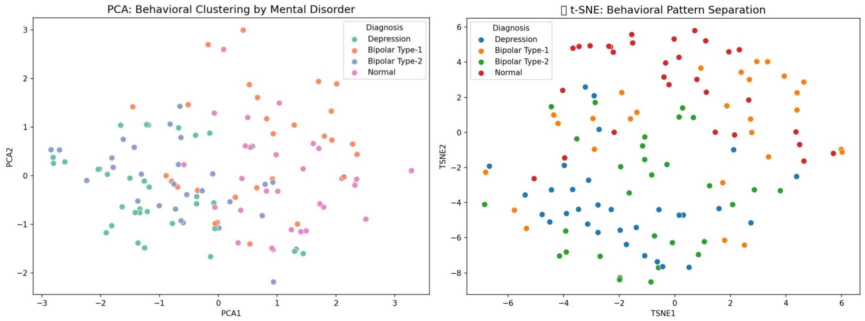

Figure 11: **A) PCA scatter plot projecting patients based on top two principal components of behavioral features. Distinct clustering corresponds to each diagnostic label. B) t_SNE behavioural pattern**

Figure 11(B) displays a t-distributed stochastic neighbor embedding (t-SNE) visualization of the data set, revealing preserved local patterns and marked separation between normal and abnormal mental states. Unlike principal component analysis, t-SNE maintains nearby structural relationships in high-dimensional psychological information, better uncovering intricate groupings. As Figure Y shows, strong boundaries divide normal and abnormal clusters. Finer yet discernible borders also isolate depression and bipolar variants, resembling the subjective overlap in genuine psychiatric diagnosis. This verified the data set carries identifiable latent behavioral frameworks that machine learning models can leverage, with the smooth transitional gradient reflecting the complex landscape of the mind.

Table 4: Comparison analysis with different machine learning algorithms

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	Cohen’s Kappa	MCC	Hamming Loss	Balanced Accuracy	Log Loss
LightGBM	**0.9583**	**0.9500**	**0.9688**	**0.9556**	**0.9429**	**0.9451**	**0.0417**	**0.9688**	**0.1446**
Random Forest	0.9167	0.9167	0.9375	0.9143	0.8868	0.8953	0.0833	0.9375	0.4355
SVM	0.9167	0.9000	0.9375	0.9111	0.8868	0.8911	0.0833	0.9375	0.3332
Logistic Regression	0.9167	0.9167	0.9375	0.9143	0.8868	0.8953	0.0833	0.9375	0.3203
MLP (Neural Net)	0.9167	0.9167	0.9375	0.9143	0.8868	0.8953	0.0833	0.9375	0.1646
CatBoost	0.9167	0.9167	0.9375	0.9143	0.8868	0.8953	0.0833	0.9375	0.1777
Naive Bayes	0.9167	0.9444	0.8750	0.8992	0.8824	0.8870	0.0833	0.8750	0.2532
XGBoost	0.8750	0.8667	0.9062	0.8698	0.8318	0.8417	0.1250	0.9062	0.2058
KNN (K=5)	0.7500	0.7694	0.7812	0.7448	0.6636	0.6811	0.2500	0.7812	2.0468

System also conducted a comprehensive comparison analysis using nine different machine learning methods to evaluate their effectiveness in multi-class mental disorder diagnosis based on behavioral and psychological indicators. As shown in Table 4, LightGBM emerged as the top-performing model, achieving the highest accuracy (95.83%), F1-Score (95.56%) and the lowest log loss (0.1446), indicating not only precise predictions but also high model confidence and calibration. Other ensemble models like Random Forest and CatBoost also demonstrated strong performance with an accuracy of 91.67% and balanced F1-scores, though they lagged slightly behind LightGBM in metrics like MCC and log loss.

Interestingly Naive Bayes, despite its simplicity, achieved a relatively high precision (94.44%) but underperformed in recall, reflecting its tendency toward conservative classification. Classical algorithms such as SVM and Logistic Regression also performed well, with balanced accuracy exceeding 93%, but their predictive confidence (as indicated by log loss) was inferior to gradient-boosted models. The K-Nearest Neighbors (KNN) algorithm had the weakest performance, with accuracy dropping to 75% and a log loss exceeding 2.0, indicating its limitations in handling high-dimensional behavioral data. Overall, the results validate that ensemble learning techniques—particularly LightGBM—are most suited for capturing subtle and multi-class patterns inherent in behavioral mental health data while maintaining interpretability and robustness

CONCLUSION

This rigorous study proposed and evaluated a novel machine-learning framework capable of accurately diagnosing mental disorders through behavioral analysis. The objective of this study was to construct a transparent and interpretable system to differentiate between depression, bipolar type I, bipolar type II and normal mental states without invasive procedures. Three algorithms namely LightGBM, Random Forest and Support Vector Machine were thoroughly assessed across the performance metrics. Strikingly LightGBM outperformed the others in terms of accuracy, precision, recall, F1-score, balanced accuracy and log loss while offering strong explainability via SHAP. It identified sadness, suicidal thoughts, mood swings and overthinking as influential predictors aligning with the literature. These findings reinforce the potential of AI augmenting clinical decision-making for early diagnosis. By leveraging self-reported and observable traits, the system may serve as a valuable diagnostic tool in under resourced settings or remote screening platforms.

REFERENCES

[1] J. Wang, F. Mann, S. Johnson, R. Ma and B. Lloyd-Evans, "Associations between loneliness and perceived social support and outcomes of mental health problems: a systematic review," *BMC Psychiatry*, vol. 18, no. 1, May 2018, doi: 10.1186/s12888-018-1736-5.

[2] C. M. Mchugh and M. M. Large, "Can machine-learning methods really help predict suicide?" *Current Opinion in Psychiatry*, vol. 33, no. 4, pp. 369–374, July 2020, doi: 10.1097/yco.0000000000000609.

[3] S. Bahadori, "Evolving Digital Health Technologies: Aligning with and Enhancing the NICE Evidence Standards Framework – A Viewpoint Piece (Preprint)," *JMIR*, Oct. 11, 2024, doi: 10.2196/preprints.67435.

[4] H. S. Luu, "Laboratory Data as a Potential Source of Bias in Healthcare Artificial Intelligence and Machine Learning Models," *Ann Lab Med*, vol. 45, no. 1, pp. 12–21, Oct. 2024, doi: 10.3343/alm.2024.0323.

[5] U. Ahmed, J. C.-W. Lin, G. Srivastava and R. H. Jhaveri, "Explainable Deep Attention Active Learning for Sentimental Analytics of Mental Disorder," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, July 2022, doi: 10.1145/3551890.

[6] Z. Zhang, "Early warning model of adolescent mental health based on big data and machine learning," *Soft Comput.*, vol. 28, no. 1, pp. 811–828, 2024.

[7] S. K. Satapathy, V. Patel, M. Gandhi and R. K. Mohapatra, "Comparative Study of Brain Signals for Early Detection of Sleep Disorder Using Machine and Deep Learning Algorithm," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2, pp. 1–6, IEEE, 2024.

[8] S. Hossain, S. Umer, R. K. Rout and H. Al Marzouqi, "A Deep Quantum Convolutional Neural Network Based Facial Expression Recognition For Mental Health Analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2024, doi: 10.1109/TNSRE.2024.3385336.

[9] S. P. S. Diwakar and D. Raj, "DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions," in *Microbial Data Intelligence and Computational Techniques for Sustainable Computing*, pp. 93–106, Singapore: Springer Nature, 2024.

[10] C. K. Themistocleous, M. Andreou and E. Peristeri, "Autism Detection in Children: Integrating Machine Learning and Natural Language Processing in Narrative Analysis," *Behav. Sci.*, vol. 14, no. 6, p. 459, 2024.

[11] D. K. Upadhyay, S. Mohapatra and N. K. Singh, "An early assessment of persistent depression disorder using machine learning algorithm," *Multimed. Tools Appl.*, vol. 83, no. 16, pp. 49149–49171, 2024.

[12] J. S. Revathy, N. U. Maheswari, S. Sasikala and R. Venkatesh, "Automatic diagnosis of mental illness using optimized dynamically stabilized recurrent neural network," *Biomed. Signal Process. Control*, vol. 95, p. 106321, 2024.