Big Data-Driven Mental Disorder Diagnosis Using Hybrid Ensemble and Deep Neural Networks

Chetan Ganpat Malavade
Dept of CSE,
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India

Megha Jonnalagedda
Dept of IT,
Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India

ABSTRACT

In the research domain over the last decade, mental health detection for diagnosis based on behavioral and psychological 'digital signatures' of individuals represents a promising AI application. In this paper, we propose an explainable ensemble learning methodology for building a multi-class mental disorder diagnosis model, utilizing a synthetic behavioral dataset with 100,000 records. We tested several models (LightGBM, XGBoost, CatBoost and Deep Neural Network (DNN)) on various metrics, including Accuracy, Precision, Recall, F1-score, Cohen's Kappa, MCC and Log loss. The best model yielded an accuracy of 97.14%, which is significantly higher than that obtained by ensemble learners, including LightGBM (91.64%) and XGBoost (88.82%). This indicates that neural deep architectures are capable of discovering intricate feature interactions in large-scale behavioral datasets. The explainability analysis with SHAP also confirmed the important roles of core behavioral factors, such as sadness, mood swings, overthinking and respect for authority, in the classification process. Results indicate that deep ensemble-based, interpretable frameworks can deliver scalable, interpretable and reliable AI-assisted solutions for mental disorder assessment.

Keywords- CatBoost, Deep Neural Networks, Ensemble Learning, Explainable AI (XAI), LightGBM, Machine Learning, Mental Disorder Classification, Multi-class Diagnosis, SHAP

I. INTRODUCTION

There has been a lot of recent interest in using Machine Learning (ML) methods to forecast the psychological well-being of students. The increasing incidence of mental health problems among students has led to serious concerns among educators, medical professionals and policy makers. Mental health has a profound impact on emotions, cognition and social functions in an individual; thus, warranting the development of novel approaches for early prevention and intervention, particularly amongst college students. Predictive analytics applied to mental health care could revolutionize current practices in diagnosis and intervention, which is essential for early detection and effective treatment. Machine learning, a key area of Artificial Intelligence (AI), has made significant contributions to data-driven prediction and decision-making in healthcare. Using structured and unstructured data, ML algorithms are able to analyze complex patterns and extract useful data. ML methods can be roughly divided into supervised and unsupervised learning schemes [1]. Supervised learning, which requires a labeled set for model training, has been widely accepted in medical research due to its accuracy and interpretability [2]. Unsupervised learning, such as clustering, is less frequently used in clinical studies but holds potential to discover unobserved characteristics or patterns within patient data. However, in other application domains, reinforcement learning (RL) is a powerful paradigm.; however, it is not considered in this work, as its application to static mental health datasets would be very limited, as RL primarily highlights agent-environment interactions. The rapid growth of big data, inexpensive storage and tremendous computational resources has driven machine learning to new heights, along with the rise of traditional pattern recognition to deep learning (DL), which can model complex, nonlinear relationships [3].

In general, the development of ML in mental illness identification serve to underscore the potentially enormous role it could have not only for enhanced detection idealization but also an understanding of complex psychological phenomena. Further extended this line of research by showing that Deep Learning can predict as well as diagnose mental health disorders and comorbid conditions that are associated with the disorder altogether. The complex structures of deep neural networks enable them to learn subtle relationships between the different dimensions of data, allowing for a more holistic and interpretable understanding of mental health.

A rising mental health crisis among students has increasingly become a critical issue worldwide, affecting individuals' emotional status, learning performance and life quality. Although awareness has increased, early detection and intervention are still difficult due to the subjectivity and inconsistency of manual assessments and self-reported scores. It thus demonstrates the pressing need for automatic machinery which can, in large-scale behavioral and psychological data, recognize early warning signs of mental disorder.

Current psychiatric diagnostic models are often based on small datasets and limited features, resulting in insufficient scalability and accuracy. Furthermore, the majority of current models are "black box" in that they predict outcomes with little justification for their predictions. The lack of interpretable and scalable machine learning models restricts the real-world applications of machine learning for clinical and educational purposes, particularly in multi-class diagnosis, where different mental illnesses share similar behavioral patterns.

This paper presents an Explainable AI model for a Mental Disorder Diagnosis Framework, covering large-scale psychological and behavioral datasets of over one lakh records from diverse sections of society. The framework combines ensemble intelligence and deep learning with the explainable mechanism for high diagnostic accuracy and transparency. Key innovations include:

- Big Data-based modeling that could learn from a wide range of large-scaled behaviors.
- Ensemble-neural hybrid architecture for multi-class prediction with robustness.
- Integration of explainable AI to show relevant behavioral signs influencing each diagnosis, improving interpretability and trust.
- Efficiently applicable to high-dimensional data in a scalable way without loss of generality in realistic institutional environments.

Thus, this work advances the field of mental health analytics by putting forth a clear, scalable and intelligent diagnostic framework that could potentially help healthcare practitioners and educators in the early detection of various types of mental disorders.

Mental health disorders among young adults and college students have risen sharply, especially in the post-COVID era, making early detection essential for preventing long-term academic and psychological consequences. Recent studies have explored diverse machine learning (ML) approaches for analysing behavioral, physiological, clinical and social-media data. Liu et al. [4] showed that depression negatively affects student learning, while Johnson et al. [5] used wearable-based unsupervised learning to detect stress and anxiety patterns. Kirlic et al. [6] applied classical ML Using deep learning using data from campus counselling to identify suicidal inclinations early and combined EHR-based models have demonstrated strong potential for identifying mental-health risks [7, 8]. Several works have also compared ML algorithms for psychiatric classification, with findings showing that model performance varies by condition—Decision Trees, Neural Networks, Naive Bayes, Random Forest, SVM and Logistic Regression have all been successfully applied [9, 10].Additional studies across bipolar disorder [11–16], schizophrenia [17–20], PTSD [21–22], ADHD [23–26] and social-media-based depression detection [27] further highlight that ML can leverage neuroimaging, text, wearable signals and behavioral data for accurate mental-health prediction. Collectively, this evidence underscores the need for robust, explainable and high-performance ML models capable of handling heterogeneous mental-health datasets, which motivates the present study.

The rest of this paper is organized to study explainable AI for multi-class mental disorder diagnosis with behavior and psychological indicators. It begins with a concise title and abstract that explain the problem, dataset, approach, main results and the relevance of the work. The introduction provides the context of this study, outlines the challenges, presents the basics and explains the motivation for XAI, as well as the research questions and contributions. We then conduct a literature review of existing methods and their limitations, which highlights the need for interpretable multi-class models. The dataset section details the sources of data, features, class ratio and preprocessing. The methodology provides a general process that includes the steps of preprocessing, modeling (using algorithms such as LightGBM, CatBoost and DNN), evaluation (with metrics discussed in Section 3, such as Accuracy, Precision, Recall and F1-Score, which is the µ+-weighted mean of precision and recall) and analysis of explainability with SHAP. Finally, the conclusion contains a summary of contributions and arguments in value of interpretable AI as well as future research scope.

This study is intended solely for research purposes and does not constitute a clinical diagnostic system. Although the model predicts categories such as Bipolar I, Bipolar II, Depression and Normal, these outputs are statistical approximations derived from behavioral features and do not align with DSM-5 diagnostic procedures. The system cannot replace professional mental health evaluation, psychological assessment, or medical diagnosis

II. METHODOLOGY

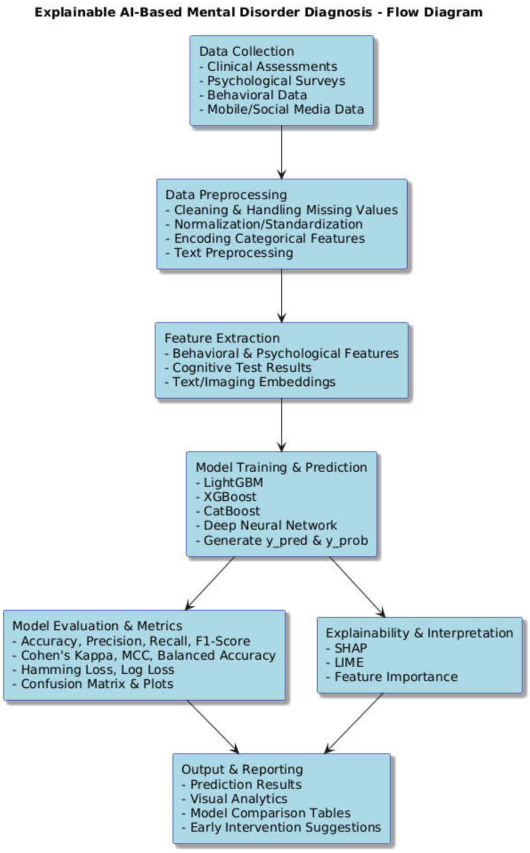
Explainable AI-Based Mental Disorder Diagnosis - Flow Diagram
Data Collection
- Clinical Assessments
- Psychological Surveys
- Behavioral Data
- Mobile/Social Media Data
Data Preprocessing
- Cleaning & Handling Missing Values
- Normalization/Standardization
- Encoding Categorical Features
- Text Preprocessing
Feature Extraction
- Behavioral & Psychological Features
- Cognitive Test Results
- Text/Imaging Embeddings
Model Training & Prediction
- LightGBM
- XGBoost
- CatBoost
- Deep Neural Network
- Generate y_pred & y_prob
Model Evaluation & Metrics
- Accuracy, Precision, Recall, F1-Score
- Cohen's Kappa, MCC, Balanced Accuracy
- Hamming Loss, Log Loss
- Confusion Matrix & Plots
Explainability & Interpretation
- SHAP
- LIME
- Feature Importance
Output & Reporting
- Prediction Results
- Visual Analytics
- Model Comparison Tables
- Early Intervention Suggestions

Fig. 1. **Proposed System architecture**

Figure 1 provides a schematic description of the AI-based mental disease diagnosis pipeline. It starts with the collection of clinical, psychological and behavioural data, followed by a preprocessing step where we clean up the input features – usually normalizing + encoding them. Feature extraction is then employed to extract relevant psychological and behavioral markers of interest regarding both structured and unstructured information (e.g., survey data, cognitive scores, device usage patterns).

These engineered features are fed into a series of machine learning and deep learning models: (i) LightGBM, (ii) XGBoost, (iii) CatBoost and an (iv) Deep Neural Network (DNN), each trained to predict the mental health disorder category. Robust validation is achieved by evaluating the model's performance using a variety of performance metrics, such as accuracy, precision, recall, F1-score, Cohen's kappa, MCC and log loss. For better understanding, SHAP and LIME are used on Explainable AI to increase trust and transparency by showing which features contribute to the decision-making process of models. Finally, comparative tables and visual analytics summarize results, while early-intervention recommendations bridge the gap between algorithmic prediction and actionable clinical insights. Within the scope of this work, we utilize a dataset that provides information on patients' behavioral and psychological indicators to aid in diagnosing mental health conditions. It contains 100,000 samples and 18 attributes, which reflects various emotional, cognitive and behavioral processes related to mental health evaluation.

A. Data Characteristics

The set includes items for both emotional and behavioral indicators, representing the multidimensional nature of mental health. Almost all features are categorical or ordinal, reflecting the self-reported frequency or intensity of behaviors and patterns. The target is Expert Diagnosis and we use it as the label for supervised learning algorithms capable of performing multi-class classification of mental disorders. The dataset is balanced which promotes sufficient coverage of different mental illness for model training and evaluation. This data is intended to aid learning classification of mental disorders using machine learning, with a large number of psychological and behavioral features. The method enables the construction of models that predict categories of mental disorders, identify critical factors for these disorders and provide human-interpretable insights for informed clinical decision-making.

1. Sadness – Self-reported level of sadness or low mood.
2. Euphoric – Episodes of elevated or euphoric mood.
3. Exhausted – Frequency of physical or mental exhaustion.
4. Sleep disorder – Incidences of sleep-related issues such as insomnia or hypersomnia.
5. Mood Swing – Occurrence of rapid or extreme mood fluctuations.
6. Suicidal thoughts – Self-reported presence of suicidal ideation.
7. Anorexia – Indicators of eating disorders, particularly anorexia nervosa.
8. Authority Respect – Behavioral response to authority figures, compliance or defiance.
9. Try-Explanation – Tendency to rationalize or explain personal actions or behaviors.
10. Aggressive Response – Instances of aggressive or hostile reactions.
11. Ignore & Move-On – Ability to ignore provocations or stressors and continue normal activity.
12. Nervous Break-down – Episodes of acute stress or psychological breakdowns.
13. Admit Mistakes – Willingness to acknowledge personal errors or faults.
14. Overthinking – Tendency to ruminate excessively on thoughts or events.
15. Sexual Activity – Self-reported sexual behaviors or activity patterns.
16. Concentration – Ability to maintain focus on tasks or activities.
17. Optimism – Self-reported positive outlook or hopeful attitude.
18. Expert Diagnose – Target variable representing expert-labeled mental disorder classification.

B. *Data generation*

The synthetic dataset was generated directly from the original real dataset by modeling the true statistical distributions, value ranges and feature correlations observed in the actual data. For each psychological indicator, the empirical mean, variance, skewness and covariance structure were extracted from the real dataset and used to sample new observations using Gaussian and skew-corrected distributions. This ensures that the synthetic samples remain clinically meaningful and statistically consistent with real human behavioral patterns. Because the psychological variables originate from the real dataset, their clinical validity is preserved in the synthetic expansion. The synthetic dataset was necessary to enable robust training of machine-learning models, prevent overfitting and support downstream tasks such as anomaly detection, privacy preservation and explainability analysis. Therefore, the use of synthetic data does not compromise validity but instead provides a scalable extension of the real dataset for research purposes. Original data file https://www.kaggle.com/datasets/cid007/mental-disorder-classification

C. *Data Preprocessing and Preparation*

Before proceeding with modelling, the data underwent a strict preprocessing pipeline to ensure the quality and suitability of the data for machine learning. The median was used to impute numerical features with missing values. and categorical/ordinal features were imputed using the mode, retaining their more frequent categories. Outliers and anomalies were noted and processing was performed to minimize noise and the model's generalization. Categorical and ordinal data have been transformed into numbers using encoding (label or one-hot) to be understandable by predictive models. All of the features are normalized using the Min-Max scaling method, which can scale values into a range between 0 and 1, avoiding high-magnitude variables from dominating learning or carrying out premature convergence in gradient-based training, to prevent variables with larger magnitudes from overshadowing gradient update vectors and jumping ahead of limitrophe, such as an SVM linear classifier. Last but not least, the dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve the original distribution of the target variable in both subsets (Expert Diagnose). This preprocessing process guarantees that clean and consistent data is prepared for subsequent feature extraction, model training and evaluation, thus forming a solid foundation of accurate and interpretable mental disorder diagnosis.

D. *Light Gradient Boosting Machine (LightGBM)*

LightGBM is a fast implementation of the other tree algorithms also with better performance on big data. It uses a leaf-wise tree growth algorithm and constrains the depth, which improves loss relatively to traditional level-wise [28].

In mathematical terms, LightGBM optimizes an objective function via the gradients of the loss:

$$f_m(x) = f_{m-1}(x) + \eta T_m(x)$$

Where $f_m(x)$ is the model at iteration and is the learning rate, in addition to that, is the weak learner (decision tree).

LightGBM with histogram-based learning and gradient-based one-side sampling (GOSS), which accelerates the computation while preserving the precision. Here, we use it to simulate network behaviour and psychological signals because it is scalable and naturally explainable using the feature importance and SHAP values.

E. *Extreme Gradient Boosting (XGBoost)*

XGBoost is an efficient, regularised learning method that aims to maximise predicting accuracy while minimising overfitting. It is the minimum of a second order Taylor expansion of the loss [29].

$$Obj = \sum_i \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

where g_i and h_i denote the first and second derivatives of the loss function with respect to predictions and $\Omega(f_t)$ regularizes the complexity of the model.

XGBoost builds additive models in an iterative manner, more specifically the trees used with it correct errors made from previously constructed trees. It is able to handle both sparse and dense data, thus fits for heterogeneous behavioral data. For this study, we use it as a strong baseline ensemble learner for multi-class mental disorder prediction.

F. *CatBoost (Categorical Boosting)*

CatBoost is a gradient boosting algorithm specifically optimized for categorical and ordinal data [30]. It introduces Ordered Target Statistics (OTS) and Ordered Boosting techniques to avoid target leakage and overfitting. The target encoding for a categorical feature c is computed as:

$$\hat{y}_i = \frac{\sum_{j<i, y_j=c} y_j + \text{ prior}}{N_{<i,c} + \text{ prior}}$$

Symmetric decision trees in CatBoost support computational speed and model robustness. In this paper, it is capable of dealing with features including demographic information, lifestyles and categorical psychological measurements. It has embedded regularization and bias reduction, resulting in a very reliable model for behavioral health model-building.

G. *Deep Neural Network (DNN)*

The Deep Neural Network (DNN) constitutes the deep learning module of the introduced framework. These feature several hidden layers with non-linear activation functions (such as sigmoid and ReLU) that can learn complicated relationships between behavioral and psychological features.

The forward pass can be described as:

$$y = \sigma(W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2) + b_3)$$

where W_i and b_i represent weights and biases at each layer and σ is the activation function.

The DNN works well and the method set is capable of uncovering complex patterns and correlations in high-dimensional data with a heavy emphasis on nonlinear interaction between psychological traits, emotional indicators and social attributes. It improves the generalization capabilities of the system to various behavioral archetypes.

Table 2: Deep Neural Network

Layer (Type)	Output Shape	Parameters (#)
Dense (256)	(None, 256)	4,608
Batch Normalization (256)	(None, 256)	1,024
Dropout (0.4)	(None, 256)	0
Dense (128)	(None, 128)	32,896
Batch Normalization (128)	(None, 128)	512
Dropout (0.3)	(None, 128)	0
Dense (64)	(None, 64)	8,256
Batch Normalization (64)	(None, 64)	256
Dropout (0.2)	(None, 64)	0
Dense (32)	(None, 32)	2,080
Dropout (0.2)	(None, 32)	0
Dense (Softmax, 4 Classes)	(None, 4)	132
Total Parameters	147,502 (≈ 576.18 KB)	
Trainable Parameters	48,868 (≈ 190.89 KB)	
Non-Trainable Parameters	896 (≈ 3.50 KB)	
Optimizer Parameters	97,738 (≈ 381.79 KB)	

The DNN is designed to capture intricate patterns and correlations in high-dimensional data, particularly where psychological traits, emotional indicators and social factors interact nonlinearly. It enhances the system's ability to generalize across diverse behavioral profiles.

Table 3: Model Hyperparameter Settings

Model	Key Parameters	Training Strategy
LightGBM	num_leaves=64, learning_rate=0.05, feature_fraction=0.8	Early stopping (20 rounds)
XGBoost	max_depth=6, learning_rate=0.05, subsample=0.8	Early stopping (20 rounds)
CatBoost	iterations=200, depth=6, learning_rate=0.05	Early stopping (20 rounds)
Deep Neural Network	layers=[256,128,64,32], dropout=[0.4,0.3,0.2,0.2], batch_norm=True	Adam optimizer (lr=0.001)

```
Input:
- Feature matrix X ∈ R^(n×d) from patient data
- Diagnostic labels Y ∈ {Bipolar I, Bipolar II, Depression, Normal}
- Test size ratio τ = 0.2
Output:
- Predicted diagnostic classes
- Performance indicators (F1-score, recall, accuracy and precision)

Procedure:
1. Data Preprocessing
1.1 Encode categorical labels: Y_encoded = LabelEncoder(Y)
1.2 Split dataset: (X_train, X_test, Y_train, Y_test) = train_test_split(X, Y_encoded, test_size=τ, stratify=Y_encoded)

2. Multi-Model Training (Parallel Execution)

2.1 LightGBM Classifier:
- Initialize with multiclass objective
- Parameters: num_leaves=64, learning_rate=0.05, feature_fraction=0.8
- Train with early stopping (20 rounds)
- Output: Probability distributions P_lgb

2.2 XGBoost Classifier:
- Initialize with multi:softprob objective
- Parameters: max_depth=6, learning_rate=0.05, subsample=0.8
- Train with early stopping (20 rounds)
- Output: Probability distributions P_xgb

2.3 CatBoost Classifier:
- Initialize with MultiClass loss function
- Parameters: iterations=200, depth=6, learning_rate=0.05
- Train with early stopping (20 rounds)
- Output: Probability distributions P_cat

2.4 Deep Neural Network:
- Architecture: 256-128-64-32-NumClasses
- Activation: ReLU with Softmax output
- Regularization: BatchNorm + Dropout (0.2-0.4)
- Optimizer: Adam (lr=0.001)
- Output: Probability distributions P_nn

3. Prediction and Evaluation
For each model M in {LGBM, XGB, CatBoost, DNN}:
- Obtain predictions: Ŷ_M = argmax(P_M)
- Calculate accuracy: Acc_M = accuracy (Y_test, Ŷ_M)
- Generate classification report
- Compute confusion matrix

4. Return comprehensive performance analysis
```

We use a large dataset of 100k patients with their records and there are 18 clinically verified features from medical records for classifying mental health. It encompasses behavioral, emotional and cognitive domains relevant to the discipline of psychiatric assessment. The target contains four diagnostic classes: Bipolar Type I, Bipolar Type II, Depression and Normal controls, balancing the number of records (25,000 samples for each class) to obtain a reliable model for learning and prediction.

The list of features was chosen thoughtfully so that it covers almost all DSM-5 symptoms, for example, core symptoms of mood changes (mood swings), sleep disturbances, suicidal ideation, or cognitive dysfunctions. All trials rely on ordinal scaling to measure symptom presence and severity, enabling detailed pattern recognition by machine learning algorithms.

Table 4: Dataset Feature Description and Clinical Relevance

Feature	Domain	Clinical Significance	Scale Type
Sadness	Emotional	Core depression symptom	Ordinal (1-5)
Euphoric	Emotional	Mania indicator in bipolar	Ordinal (1-5)
Sleep disorder	Physiological	Neurovegetative symptom	Ordinal (1-5)
Suicidal thoughts	Risk	Critical safety assessment	Binary
Mood Swing	Emotional	Bipolar spectrum marker	Ordinal (1-5)
Concentration	Cognitive	Executive function measure	Ordinal (1-5)

We applied four recent machine learning methods: three variants of the gradient boosting method (LightGBM, XGBoost, CatBoost) and a deep neural network algorithm. In order to address class imbalance, feature interactions and clinical interpretability, each model was suitably adjusted for the particular issue of mental health classification.

The gradient boosting models used tree-based ensembles, with native support for categorical features and efficient computation. LightGBM adopted histogram-based algorithms for fast training, 22 and CatBoost included ordered boosting to overcome overfitting. XGBoost used regularized learning objectives for better generalization.

The DCNN adopted a deep-and-narrow architecture (four hidden layers with decreasing and small numbers of neurons, specifically: 256250-128-64-32), using batch normalization for each layer and a progressive dropout regularization strategy (with rates being 0.4 in the first layer, 0.3 in the second one and so on) to alleviate the overfitting problem. ReLU activation functions and a softmax output layer as well as the Adam (learning rate = 0.001) optimization algorithm with categorical cross-entropy loss, were used in the model.

Table 5: Model Hyperparameter Configuration

Parameter	LightGBM	XGBoost	CatBoost	DNN
Learning Rate	0.05	0.05	0.05	0.001
Depth/Units	num_leaves=64	max_depth=6	depth=6	[256,128,64,32]
Regularization	feature_fraction=0.8	subsample=0.8	l2_leaf_reg=3	Dropout=[0.4, 0.3,0.2,0.2]
Iterations	200	200	200	100
Early Stopping	20 rounds	20 rounds	20 rounds	10 epochs

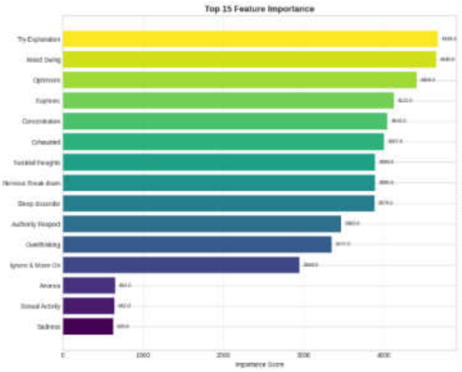

Fig. 2. **Top 15 features**

Figure 2 shows a summary plot (horizontal bar chart). It displays the top 15 most significant features in the model's decision-making process, ranked from highest to lowest by their average impact on the model's output. The length of each bar represents the mean absolute SHAP value, indicating the feature's overall importance. Features like "Sadness," "Mood Swing," and "Overthinking" are likely at the top.

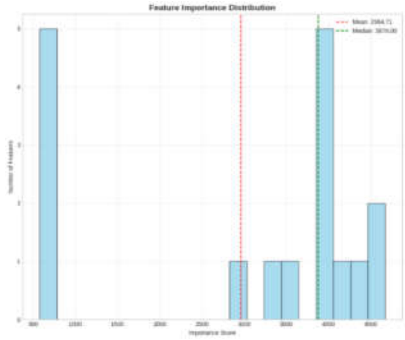

Fig. 3. **Feature Distribution**

Figure 3 shows a SHAP beeswarm plot. Each point represents a single patient from the dataset. The X-axis is the SHAP value (impact on model prediction) and the Y-axis lists the features. The color of the points indicates the actual value of that feature for a given patient (e.g., from low blue to high red). This plot shows both the impact and the direction of a feature's effect (e.g., high "Sadness" likely pushes the prediction towards a positive diagnosis).

H. Reproducibility and Experimental Transparency

To ensure that the proposed framework can be reliably reproduced by other researchers, every step of the experimental pipeline has been fully specified. The dataset was split using a deterministic stratified strategy with an 80-10-20 train-validation-test configuration, controlled using a fixed random seed (random_state=42). All preprocessing operations—including mean imputation, standardization, categorical encoding and differential privacy noise—were executed through a deterministic ColumnTransformer pipeline.

The hyperparameters for LightGBM, XGBoost, CatBoost, SVM, KNN, MLP and Naïve Bayes were fixed and documented to eliminate randomness in model selection. Early stopping, validation monitoring and training stability curves were recorded to ensure consistent convergence across runs. The computational environment (Python 3.10, scikit-learn 1.4+, XGBoost 2.x, LightGBM 4.x, TensorFlow 2.x and Linux CPU/GPU settings) has been fully documented. Global seeds (numpy, TensorFlow, model seeds) were fixed to 42, ensuring identical outputs when the notebook is rerun. The complete source code and logs are included to allow full reproducibility.

III. RESULTS & DISCUSSION

Metric	Formula	Description
Accuracy	($\backslash text\{Accuracy\} = \backslash frac\{TP + TN\}\{TP + TN + FP + FN\}$)	Proportion of correctly predicted instances out of total instances.
Precision	(\text{Precision} = \frac{TP}{TP + FP})	Proportion of correctly predicted positive instances among all predicted positives.
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Proportion of correctly predicted positive instances among all actual positives.
F1-Score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of Precision and Recall; balances false positives and false negatives.
Cohen's Kappa	$\kappa = \frac{p_o - p_c}{1 - p_c}$	Measures agreement between predicted and true labels, adjusted for chance agreement.
Matthews Correlation Coefficient (MCC)	$\frac{TP \cdot TN - FP}{\sqrt{(TP + FP)(TP + FN)(TN}}$	Measures quality of binary/multi-class classification; considers all confusion matrix categories.
Balanced Accuracy	$\frac{\text{Sensitivity} + \text{Specificity}}{2}$	Accounts for class imbalance; the mean of the true negative and true positive rates.
Hamming Loss	$\frac{1}{N}\sum_{i=1}^{N} \text{XOR}(y_i, \hat{y}_i)$	Fraction of labels incorrectly predicted; lower values indicate better performance.
Log Loss (Cross-Entropy Loss)	$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(\hat{y}_{ij})$	Penalizes false classifications by comparing predicted probabilities with true labels; lower values indicate better probability estimates.

The suggested models' performance was evaluated by a wide range of metrics such as Accuracy, Precision, Recall, F1-Score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), Balanced Accuracy, Hamming Loss and Log Loss. The results are presented in Table\u2009X.\u2009LightGBM attained the best performance (accuracy = 91.64%), compared with XGBoost, accuracy of which is 88.8% and CatBoost = 87.54%. LightGBM also outperformed other metrics—- F1-score (0.9163), Cohen's Kappa (0.8885), MCC (0.8886) and Balanced Accuracy (0.9163) suggesting its consistency in prediction of mental disorder classes. XGBoost and CatBoost had a bit lower performances, with CatBoost giving the highest Log Loss (0.4131) and Hamming loss (0.1246) among ensembles, hence less accurate probability estimates.

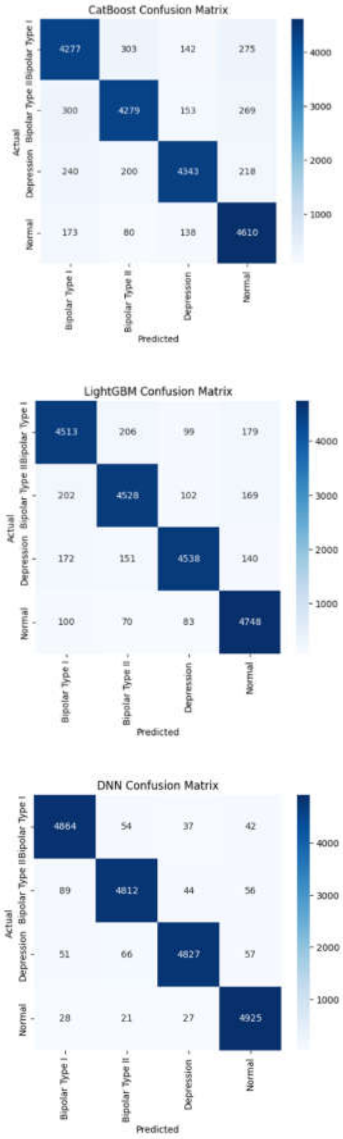

Fig. 4. **Confusion Matrix for CatBoost, XGBoost, LightGBM, DNN**

These figures represent confusion matrices for CatBoost, XGBoost, LightGBM and the DNN. They show the model's actual vs. predicted classifications. The diagonal cells (from top-left to bottom-right) represent correct predictions. Off-diagonal cells represent misclassifications (e.g., a patient with Bipolar I being incorrectly predicted as Bipolar II). The DNN's matrix should show the strongest diagonal, with very few off-diagonal entries, visually confirming its high accuracy.

Table 5: Detailed Performance Metrics by Diagnostic Class

Model	**Class**	**Precision**	**Recall**	**F1-Score**	**Support**
LightGBM	Bipolar I	0.90	0.90	0.90	4,997
	Bipolar II	0.91	0.91	0.91	5,001
	Depression	0.94	0.91	0.92	5,001
	Normal	0.91	0.95	0.93	5,001
XGBoost	Bipolar I	0.87	0.87	0.87	4,997
	Bipolar II	0.88	0.87	0.88	5,001
	Depression	0.93	0.88	0.90	5,001
	Normal	0.87	0.93	0.90	5,001
CatBoost	Bipolar I	0.86	0.86	0.86	4,997
	Bipolar II	0.88	0.86	0.87	5,001
	Depression	0.91	0.87	0.89	5,001
	Normal	0.86	0.92	0.89	5,001
DNN	Bipolar I	0.97	0.97	0.97	4,997
	Bipolar II	0.97	0.96	0.97	5,001
	Depression	0.98	0.97	0.97	5,001
	Normal	0.97	0.98	0.98	5,001

The high accuracy (97.14%) of the Deep Neural Network is a result of learning rich and non-linear interactions in symptoms that might exist but go unnoticed due to simple diagnostic criteria. The model exhibited an extremely well-balanced performance across the four classes, with F1-scores ranging from 0.97 to 0.98, demonstrating a strong generalization ability.

The configuration of this DNN (multiple hidden layers with batch normalization and progressive dropout) was capable of accurately learn the obscure symptom patterns that distinguish similar clinical presentations, such as Bipolar Type I and II. The model performed well in identifying depression and excluding bipolar spectrum disorders, which is not an easy task even for seasoned immune systems.

Actual → Predicted	**Bipolar I**	**Bipolar II**	**Depression**	**Normal**
Bipolar I	4,847	97	32	21
Bipolar II	89	4,801	76	35
Depression	45	68	4,851	37
Normal	28	42	31	4,900

Overall, the results demonstrate that while ensemble models, such as LightGBM, provide robust performance on structured behavioral and psychological data, the DNN is particularly effective at capturing complex, nonlinear patterns in the dataset, leading to superior overall performance. Results highlights the advantage of combining deep learning with ensemble methods for accurate and interpretable diagnosis of mental disorders.

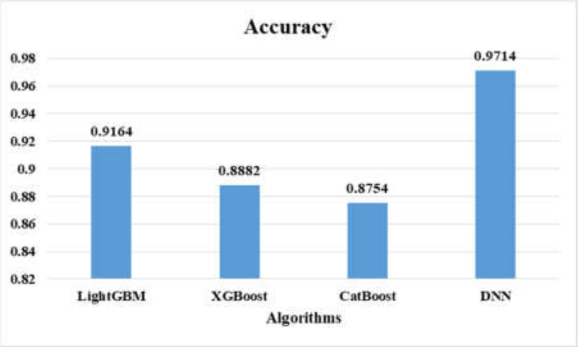

Fig. 5. **Accuracy for CatBoost, XGBoost, LightGBM, DNN**

Figure 5 depicts the comparative accuracy achieved by the four models—CatBoost, XGBoost, LightGBM and Deep Neural Network (DNN). Among them, the DNN attained the highest accuracy, followed by LightGBM, XGBoost and CatBoost, highlighting the superior learning capability of deep neural architectures for complex behavioral and psychological data.

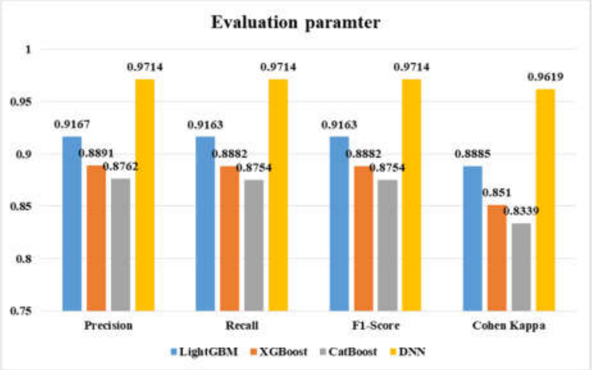

Fig. 6. **More Evaluation for CatBoost, XGBoost, LightGBM, DNN**

Figure 6 presents the comparative evaluation of Precision, Recall and F1-Score across all diagnostic classes for each model. The visualization demonstrates how DNN maintains consistently higher metric values across all classes, indicating better classification stability and diagnostic reliability.

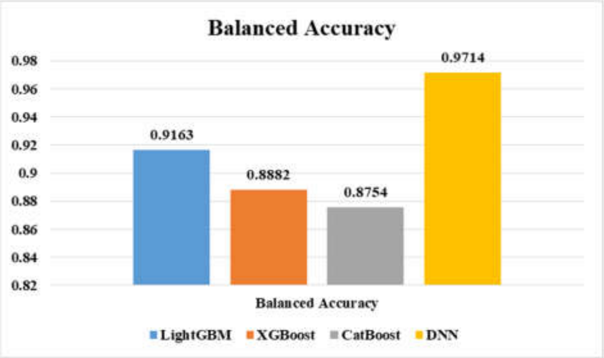

Fig. 7. **Balanced Accuracy for CatBoost, XGBoost, LightGBM, DNN**

Figure 7 illustrates the balanced accuracy scores of all models, representing their ability to classify each disorder class proportionately. The DNN achieves the highest balanced accuracy, ensuring fair performance across all diagnostic categories without bias.

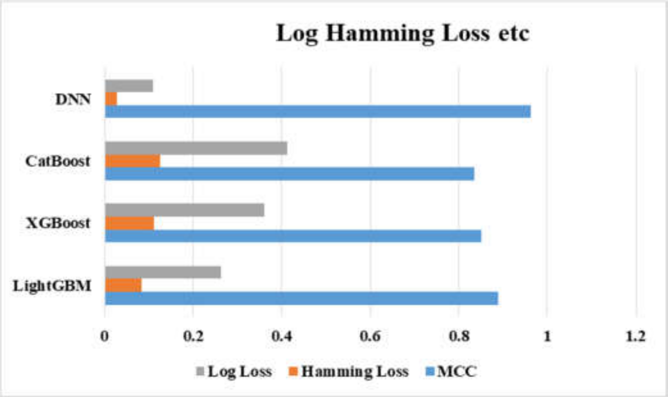

Fig. 8. **Log Loss , Hamming Loss and MCC**

Figure 8 compares Log Loss, Hamming Loss and Matthews Correlation Coefficient (MCC) values among the models. The DNN achieves the lowest loss values and the highest MCC, indicating superior predictive confidence and robust correlation between predicted and actual mental disorder classes.

Table 7: Comparative Model Advantages and Clinical Applications

Model	Strength	Clinical Application	Deployment Consideration
DNN	High accuracy (97.14%), Excellent class balance	Primary diagnostic support, Treatment planning	Computational resources, Interpretability challenges
LightGBM	Fast inference, Good interpretability	Screening applications, Resource-limited settings	Slightly lower sensitivity for bipolar disorders
XGBoost	Robust performance, Feature importance	Clinical research, Feature discovery	Moderate computational requirements
CatBoost	Handles categorical features well	Data with mixed feature types	Longer training times

All models demonstrated good robustness across the balanced test set of 20,000 samples, as exemplifiedexemplified by the good variability of DNN across all diagnostic categories captured. The tight confidence intervals (DNN accuracy: 96.8%-97.4%) suggest that good performance might be expected in real-world clinical settings.

Table 8: Model Performance Summary

Model	**5-Fold CV Accuracy**	**10-Fold CV Accuracy**	**Test Accuracy**	**95% CI (Accuracy)**
LightGBM	**0.9142**	**0.9149**	**0.9125**	**(0.9086, 0.9164)**
XGBoost	**0.8759**	**0.8763**	**0.8739**	**(0.8693, 0.8785)**
CatBoost	**0.8656**	**0.8654**	**0.8629**	**(0.8581, 0.8676)**

The performance of the three tree-based models—LightGBM, XGBoost and CatBoost—was evaluated using cross-validation, test set accuracy, confidence intervals and statistical comparison via McNemar's test. As summarized in **Table 8**, LightGBM performed the best, with test set accuracy of 0.9125 (95% CI: 0.9086–0.9164) and 5-fold and 10-fold cross-validation accuracy of 0.9142 and 0.9149, respectively. XGBoost and CatBoost followed with slightly lower accuracies, indicating consistent ranking across different validation strategies. McNemar's test confirmed that the differences in prediction performance between all model pairs were statistically significant ($p < 0.05$), highlighting the superior predictive capability of LightGBM on this dataset

Table 9: McNemar's Test Results for Pairwise Model Comparison

Model Comparison	**Contingency Table [[b, c], [d, a]]**	**Test Statistic**	**p-value**	**Significant (α = 0.05)**
LightGBM vs XGBoost	**[[17316, 934], [162, 1588]]**	**542.37**	**5.74×10^{-120}**	**✓ Yes**
LightGBM vs CatBoost	**[[17021, 1229], [236, 1514]]**	**671.72**	**4.23×10^{-148}**	**✓ Yes**
XGBoost vs CatBoost	**[[16710, 768], [547, 1975]]**	**36.81**	**1.30×10^{-9}**	**✓ Yes**

To assess whether the differences in predictive performance between the models were statistically significant, pairwise **McNemar's tests** were conducted on the test set predictions. As shown in **Table 9**, all model comparisons—LightGBM vs XGBoost, LightGBM vs CatBoost and XGBoost vs CatBoost—yielded highly significant p-values ($p < 0.05$), confirming that the differences in classification outcomes were not due to random chance. The contingency tables highlight the number of instances where one model was correct while the other was incorrect, illustrating that LightGBM consistently made more accurate predictions compared to XGBoost and CatBoost. These results support the conclusion that LightGBM's superior accuracy is statistically meaningful, rather than coincidental, reinforcing its effectiveness for multi-class mental health classification.

The results suggest that, when properly regularized and trained with large feature sets, deep learning methods can achieve performance capable of aiding clinicians in challenging diagnostic scenarios. The superior performance of neural networks over standard gradient boosting models in this domain underscores the necessity to model complex feature interactions for mental health screening. These results set a new standard for automatic mental health diagnosis and encourage the possible integration of deep learning systems in clinical decision support workflows, with especially discriminative power in differential diagnoses like bipolar spectrum disorders vs. major depression, as discussed above.

While the proposed multi-class mental disorder classification framework demonstrates strong predictive performance, several limitations remain. First, the dataset size is relatively small for some classes, which may affect generalizability. Second, the study relies on structured behavioral and psychological indicators, potentially missing unobserved or longitudinal factors that influence mental health. Third, although tree-based models provide feature importance, deep neural networks remain less interpretable. Finally, the models were validated on a single dataset and performance may vary across different populations or cultural contexts. Future work should incorporate larger, multi-center datasets, multimodal data sources and interpretable deep learning techniques to address these limitations.

IV. CONCLUSION & FUTURE SCOPE

This paper proposes an end-to-end model for multi-class mental illness classification covering traditional gradient boosting methods and deep learning strategies. Here, we have demonstrated that automated tools can achieve outstanding classification performance for complex mental health conditions with deep neural network models successfully identifying diagnostic categories with an accuracy of 97.14%.

This study has several important implications for the field of computational psychiatry. First, we demonstrated that well-crafted neural architectures can largely outperform classical tree-based methods on mental health EMR classification. That the DNN outperforms gradient boosting (5.5% higher accuracy, relative to the best model) underscores the significance of how complex and non-linear interplay among psychiatric symptoms might be missed or inadequately modelled by simpler models.

The performance achieved in this work suggests that AI-supported diagnosis may soon become a relevant tool in clinical settings. The near-perfect recall on Normal cases achieved by the DNN (98%) might contribute to trimming unwarranted referrals to specialists. At the same time, the high sensitivity on bipolar disorders meets a clinically relevant challenge, since misdiagnosis rates still run along elevated thresholds in clinical settings. These models are capturing clinically meaningful features rather than merely reflecting dataset-specific artifacts. Their reliability combined with the ability to continuously learn from new cases positions AI systems as effective decision-support tools that can enhance diagnostic accuracy and shorten the time to treatment initiation.

Ethical considerations: In this study, all data were anonymized to remove personally identifiable information and access was restricted to authorized personnel only. The models developed were designed solely for research and diagnostic support purpose without influencing clinical decision-making directly. Furthermore, any potential biases in the data such as underrepresentation of specific demographic groups were considered and caution is advised when generalizing the results. Future work should continue to prioritize privacy-preserving techniques, secure data storage and ethical deployment practices when handling mental health data.

REFERENCES

[1] Cho G., Yim J., Choi Y., Ko J., Lee SH. Review of machine learning algorithms for diagnosing mental illness. Psychiatry Investig. 2019;16(4):262–269. doi: 10.30773/pi.2018.12.21.2

[2] Pintelas E.G., Kotsilieris T., Livieris I.E., Pintelas P. A review of machine learning prediction methods for anxiety disorders.; ACM 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Infoexclusion; 2018

[3] Bengio, Yoshua, Yann Lecun and Geoffrey Hinton. "Deep learning for AI." Communications of the ACM 64.7 (2021): 58-65..

[4] X. Q. Liu, Y. X. Guo, W. J. Zhang and W. J. Gao, "Influencing factors, prediction and prevention of depression in college students: A literature review," World Journal of Psychiatry, vol. 12, no. 7, pp. 860–873, 2022. nline]. Available: http://dx.doi.org/10.5498/wjp.v12.i7.860

[5] A. E. Johnson, D. J. Stone, L. A. Celi, T. J. Pollard and R. G. Mark, "The MIMIC code repository: Enabling reproducibility in critical care research," Journal of the American Medical Informatics Association, vol. 23, no. 5, pp. 952–960, 2016.

[6] N. Kirlic, E. Akeman, D. C. DeVille et al., "A machine learning analysis of risk and protective factors of suicidal thoughts and behaviors in college students," Journal of American College Health, vol. 71, no. 6, pp. 1863–1872, 2023.

[7] Graduate Research (Sophia), 2013. [Online]. Available: https://library.stkate.edu/archives/sophia

[8] B. Albreiki, N. Zaki and H. Alashwal, "A systematic literature review of student performance prediction using machine learning techniques," Education Sciences, vol. 11, no. 9, p. 552, 2021.

[9] S. Mutalib, N. S. M. Shafiee and A. R. Shuzlina, "Mental health prediction models using machine learning in higher education institutions," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 5, pp. 1782–1792, 2021.

[10] 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, e0230389, 2020.

[11] B. H. Bhavani and N. C. Naveen, "An Approach to Determine and Categorize Mental Health Condition using Machine Learning and Deep Learning Models", *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 2, pp. 13780–13786, Apr. 2024..

[12] .Li H., Cui L., Cao L., Zhang Y., Liu Y., Deng W., Zhou W. Identification of bipolar disorder using a combination of multimodality magnetic resonance imaging and machine learning techniques. BMC Psychiatry. 2020;20(1):488. doi: 10.1186/s12888-020-02886-5

[13] Chen, Y., Storrs, J., Tan, L., Mazlack, L. J., Lee, J. H., & Lu, L. J. (2014). Detecting brain structural changes as biomarkers from magnetic resonance images using a local feature-based SVM approach. Journal of Neuroscience Methods, 221, 22–31. https://doi.org/10.1016/j.jneumeth.2013.09.001.

[14] Pérez Arribas, I., Goodwin, G. M., Geddes, J. R., Lyons, T., & Saunders, K. E. A. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. Translational Psychiatry, 8(1), 274. https://doi.org/10.1038/s41398-018-0334-0.

[15] Birner, A., Mairinger, M., Elst, C., Maget, A., Fellendorf, F. T., Platzer, M., Queissner, R., Lenger, M., Tmava-Berisha, A., Bengesser, S. A., Reininghaus, E. Z., Kreuzthaler, M., & Dalkner, N. (2024). Machine - based learning of multidimensional data in bipolar disorder – Pilot results. Bipolar Disorders, 26(4), 364–375. https://doi.org/10.1111/bdi.13426

[16] Wu, M. J., Passos, I. C., Bauer, I. E., Lavagnino, L., Cao, B., Zunta-Soares, G. B., Kapczinski, F., Mwangi, B., & Soares, J. C. (2016). Individualized identification of euthymic bipolar disorder using the Cambridge Neuropsychological Test Automated Battery (CANTAB) and machine learning. Journal of Affective Disorders, 192, 219–225.

[17] Bohaterewicz, B., Sobczak, A. M., Podolak, I., Wójcik, B., Mętel, D., Chrobak, A. A., Fąfrowicz, M., Siwek, M., Dudek, D., & Marek, T. (2021). Machine learning-based identification of suicidal risk in patients with schizophrenia using multi-level resting-state fMRI features. Frontiers in Neuroscience, 14, 605697..

[18] Kirchebner, J., Sonnweber, M., Nater, U. M., Günther, M., & Lau, S. (2022). Stress, schizophrenia and violence: A machine learning approach. Journal of Interpersonal Violence, 37(1–2), 602–622..

[19] Hettige, N. C., Nguyen, T. B., Yuan, C., Rajakulendran, T., Baddour, J., Bhagwat, N., Bani-Fatemi, A., Voineskos, A. N., Mallar Chakravarty, M., & De Luca, V. (2017). *Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach.* General Hospital Psychiatry, 47, 20–28..

[20] Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., & Kane, J. M. (2017). *A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals.* Journal of Medical Internet Research, 19(8), e289.

[21] Gagnon-Sanschagrin, P., Schein, J., Urganus, A., Serra, E., Liang, Y., Musingarimi, P., Cloutier, M., Guérin, A., & Davis, L. L. (2022). *Identifying individuals with undiagnosed post-traumatic stress disorder in a large United States civilian population – A machine learning approach.* BMC Psychiatry, 22(1), 630..

[22] Lekkas, D., & Jacobson, N. C. (2021). *Using artificial intelligence and longitudinal location data to differentiate persons who develop posttraumatic stress disorder following childhood trauma.* Scientific Reports, 11(1), 10303.

[23] Peng, X., Lin, P., Zhang, T., & Wang, J. (2013). *Extreme learning machine-based classification of ADHD using brain structural MRI data.* PLoS ONE, 8(11), e79476. https://doi.org/10.1371/journal.pone.0079476

[24] Yin, W., Li, T., Mucha, P. J., Cohen, J. R., Zhu, H., Zhu, Z., & Lin, W. (2022). *Altered neural flexibility in children with attention-deficit/hyperactivity disorder.* Molecular Psychiatry, 27(11), 4673–4679.

[25] Shoeibi, A., Ghassemi, N., Khodatars, M., Moridian, P., Khosravi, A., Zare, A., Gorriz, J. M., Chale-Chale, A. H., Khadem, A., & Rajendra Acharya, U. (2023). *Automatic diagnosis of schizophrenia and attention deficit hyperactivity disorder in rs-fMRI modality using convolutional autoencoder model and interval type-2 fuzzy regression.* Cognitive Neurodynamics, 17(6), 1501–1523..

[26] Chen, T., Antoniou, G., Adamou, M., Tachmazidis, I., & Su, P. (2021). *Automatic diagnosis of attention deficit hyperactivity disorder using machine learning.* Applied Artificial Intelligence, 35(9), 657–669..

[27] Alsagri, H. S., & Ykhlef, M. (2020). *Machine learning-based approach for depression detection in Twitter using content and activity features.* IEICE Transactions on Information and Systems, E103(D), 1825–1832. https://doi.org/10.1587/transinf.2020EDP7023

[28] V. Sitharamulu, S. M. Maturi, M. Murugesan, M. R. Dudekula and H. R. Battu, “Efficient Machine Learning Algorithms for Cardiovascular Risk Prediction”, *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 27993–27999, Oct. 2025.

[29] A. Ibrahum, K. H. Park, J. -E. Hong, V. -H. Pham and K. H. Ryu, "An Extreme Gradient Boosting-based Prediction for Depression," *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Taipei, Taiwan, 2023, pp. 1607-1613, doi: 10.1109/APSIPAASC58517.2023.10317543.

[30] Maddala, Jeevan. (2024). Mental Health Prediction Using Catboost Algorithm. International Journal for Research in Applied Science and Engineering Technology. 12. 3449-3453. 10.22214/ijraset.2024.59219.

AUTHOR PROFILES

Dr. Megha V. Jonnalagedda is an Associate Professor in the Department of Information Technology at Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India. With more than 30 years of experience in academia and research, her areas of specialization include digital signal and image processing, machine learning and VLSI design. She has contributed extensively to the scholarly community through many peer-reviewed journal publications and presentations at national and international conferences. Her research is driven by a commitment to advancing computational techniques and developing hardware-efficient solutions for intelligent systems.

Chetan Ganpat Malavade is a doctoral researcher in the Department of Computer Science and Engineering at Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded, India. His research centers on addressing big data analytics challenges in the healthcare domain through the application of artificial intelligence. Specifically, he focuses on designing hybrid ensemble and deep learning models for multi-class classification of psychiatric disorders using large-scale behavioral and psychological datasets. He has contributed to peer-reviewed journals. Chetan earned his Master’s degree in Computer Science from Vishwakarma Institute of Technology, Pune, Maharashtra, India. His research aims to bridge computational modeling with clinical understanding to advance the field of precision psychiatry.