Impact Factor 6.1

Journal of Cyber Security

ISSN:2096-1146

PRIVACY-PRESERVING EXPLAINABLE AI FOR MULTI-CLASS MENTAL DISORDER DIAGNOSIS FROM BEHAVIORAL AND PSYCHOLOGICAL METRICS

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: *This research addresses the challenge of multi-class mental disorder diagnosis (using behavioral, psychological metrics) while preserving the privacy of the data and providing post hoc explanations to a set of humans. Error-prone samples were removed using deep autoencoder based global anomaly detection technique that significantly improved data quality as reflected through multiple number of high-AUC machine learning models' performances over the 100,000 records datasets. For reliable mental disorder classification, the study uses an extensive multi-model framework that integrates MLP, SVM, Random Forest, LightGBM, CatBoost, XGBoost, KNN, Naive Bayes and a Stacking Ensemble. Out of all the models trained, MLP domains with the highest metrics 96.08% accuracy, 0.9608 macro-F1, MCC 0.9477 and Cohen's Kappa 0.9477. Almost 5% of these patterns were revealed by feeding the data into an anomaly detection system. Proposed system revealed enviably high-risk metrics using SHAP. In order to create an accurate and intelligible evaluation system, this work has accomplished secure data preprocessing, anomaly filtering, multiclass mental disorder classification and SHAP explanation based on the SHAP. In addition to providing direct psychological risk exposure, the suggested integrated model ensures privacy protection, a high-quality single-cleaned sample and equitable predictive performance.*

Keywords: Mental disorder, SHAP, data privacy, structured data, predictive analysis, big data, healthcare

1 INTRODUCTION

Mental health disorders, which impact psychological well-being, cognitive functioning, emotional stability and behavioral expression, have become a major global concern [1]. Early detection of mental disorders has become a crucial research priority due to increased stress levels, lifestyle imbalances, environmental pressures and restricted access to clinical professionals. Due to their heavy reliance on subjective evaluations and clinical interviews, conventional diagnostic techniques are prone to bias, inconsistent results and intervention delays. Because AI and ML can recognise subtle symptoms, learn complex behavioral patterns and support scalable, data-driven decisions, their use in mental health diagnosis has increased dramatically.

According to behavioral psychology, there is a strong correlation between underlying mental disorders and emotions like sadness, euphoria, aggression, sleep disturbance, suicidal thoughts, mood swings and cognitive indicators like overthinking, attention, concentration and optimism. However, because of data imbalance, noise, subject variability and the sensitive nature of psychological data, modelling these multi-dimensional behavioral indicators is still difficult [2]. Anomalies, extreme outliers and privacy issues can impact traditional machine learning models, which can lower trust and limit their practicality [3,4].

The majority of current research relies on small datasets, binary classification and single-model approaches that are unable to fully capture the complex, multifaceted nature of psychological disorders, even though machine learning is increasingly being used in mental health assessment [5,6]. In addition to ignoring feature-level behavioral and psychological indicators [7] that are essential for real-world diagnosis, current systems frequently lack explainability, making it challenging for clinicians to trust or interpret model decisions. Furthermore, there is a dearth of scalable, generalizable models appropriate for multi-class diagnosis and a large gap in comparative evaluation across various algorithms using robust metrics. By combining a thorough multi-model framework, sophisticated explainability techniques and multi-dimensional behavioral features, this project fills in these gaps and creates a more accurate, comprehensible and clinically significant mental disorder diagnosis system. The current study suggests a safe, privacy-preserving, anomaly-aware AI pipeline for multi-class mental disorder diagnosis using behavioral and psychological indicators in order to address these issues. The framework consists of:

- Differential Privacy (DP) to guarantee confidentiality and protection at the attribute level.
- Autoencoder-based anomaly detection to remove anomalous behavioral samples with severe reconstruction errors.
- Thorough machine learning benchmarking, encompassing MLP, SVM, LightGBM, CatBoost, KNN, Naive Bayes, XGBoost and stacking.
- Explainable AI (XAI) utilising SHAP to interpret psychological risk factors influencing diagnostic results [8,9].

After processing a dataset of 100,000 behavioral records, about 5% of high-error anomalies were found and eliminated. These results confirm that incorporating explainability, anomaly reduction and privacy protection greatly increases the dependability and confidence of mental health prediction systems.

All things considered, the suggested framework offers an AI system for early multi-class mental disorder screening that is safe, comprehensible and clinically aligned. Large-scale psychological surveys, clinical decision support systems and digital mental health platforms could all benefit from its strong behavioral insights and high accuracy.

2 LITRAETURE REVIEW

In order to improve early psychological disorder detection, researchers are investigating deep learning, natural language processing, behavioral modelling and biomedical signal analysis. This has led to a significant increase in AI-driven mental-health analytics in recent years. A hybrid CNN–LSTM framework for early adolescent depression detection was presented by Zhang et al. [10] in 2024. Their model demonstrated the potential of multimodal deep neural architectures for clinical early-warning systems by achieving highly competitive performance with a 92% F1-score and 97% AUC using a large-scale dataset that included electronic health records and neuroimaging data from over 50,000 teenagers.

Similar to this, Satapathy et al. [11] looked into a number of machine learning and deep learning methods for categorising sleep disorders like narcolepsy, insomnia and sleep apnoea. Using EEG-based temporal patterns, their study demonstrated the efficacy of deep spatio-temporal feature extraction for neurophysiological signals, with CNN and RNN architectures outperforming traditional ML algorithms. Hossain et al. [12] presented an automated facial-expression monitoring system in a different 2024 study that combined quantum and classical deep learning models. To monitor emotional fluctuations, the framework used both static medical images and video streams. Their innovative five-stage fusion method, which combined traditional and quantum-derived decision scores, improved emotion recognition accuracy, underscoring the growing significance of quantum-enhanced AI in mental health research.

Diwakar and Raj [13] created a DistilBERT-based classification model for text-based diagnostics that can recognise mental health conditions from written user content, such as autism, anxiety and borderline personality disorder (BPD). Their model attained 96% accuracy using a balanced dataset with 500 samples per class. In keeping with an interdisciplinary trend in mental-health analytics, the study also examined preliminary evidence connecting gut microbiota patterns to neuropsychological disorders. Peristeri et al. [14] expanded on behavioral interpretation by introducing an AI-driven framework that uses XGBoost and NLP techniques to use storytelling narratives to distinguish children with Autism Spectrum Disorder (ASD) from neuro-typical children. There were 68 kids with ASD and 52 kids with typical development in the dataset. Significant behavioral and linguistic divergences were found and gradient boosting models successfully captured these differences.

Upadhyay et al. [15] also investigated behavioral data-driven prediction, using a stacking ensemble of SVM classifiers for PDD early detection. Their demographic analysis highlighted the socio behavioral aspects of disorder vulnerability by showing higher PDD prevalence among middle-class students enrolled in non-technical academic programs and among rural students across extreme income groups.

Lastly, the potential of a Dynamically Stabilised Recurrent Neural Network (DSRNN) for improved feature extraction and diagnostic accuracy in mental-health classification tasks was shown by Revathy et al. [16]. Their model demonstrated the significance of stability-aware recurrent architectures for psychological data by capturing frequency-domain relationships between healthy and affected individuals using the OSMI dataset.

3 SYSTEM METHODOLOGY

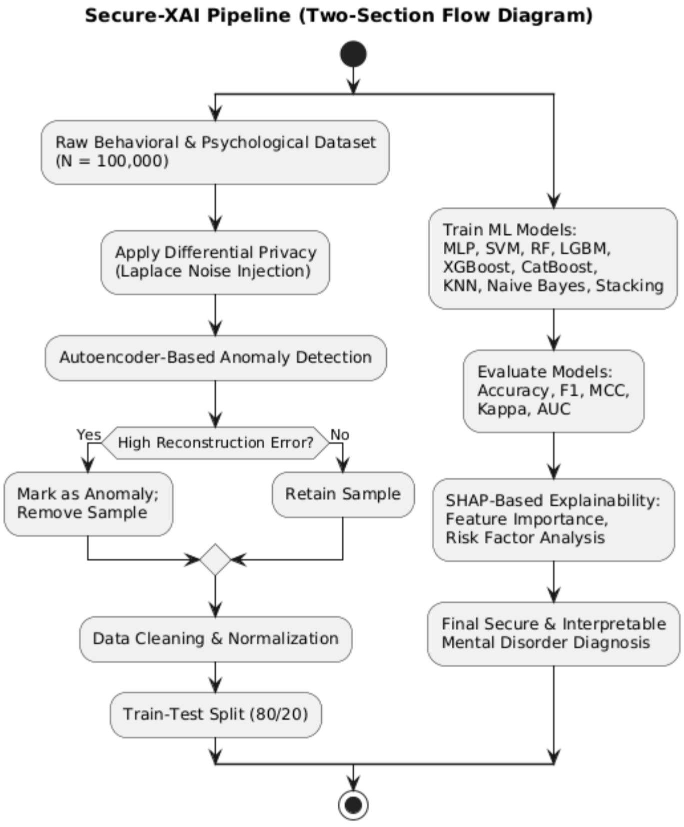
Secure-XAI Pipeline (Two-Section Flow Diagram)
Raw Behavioral & Psychological Dataset (N = 100,000)
Apply Differential Privacy (Laplace Noise Injection)
Autoencoder-Based Anomaly Detection
High Reconstruction Error?
Yes
No
Mark as Anomaly; Remove Sample
Retain Sample
Data Cleaning & Normalization
Train-Test Split (80/20)
Train ML Models: MLP, SVM, RF, LGBM, XGBoost, CatBoost, KNN, Naive Bayes, Stacking
Evaluate Models: Accuracy, F1, MCC, Kappa, AUC
SHAP-Based Explainability: Feature Importance, Risk Factor Analysis
Final Secure & Interpretable Mental Disorder Diagnosis

Figure. 1. System architecture

Figure 1 shows combination of differential privacy, anomaly detection, multi-model machine learning and SHAP-based explainability are all integrated into a single analysis pipeline in the proposed Secure-XAI Mental Disorder Diagnosis Framework. The following elements make up the methodology: (1) Data privacy protection; (2) anomaly detection; (3) pre-processing; (4) model development; (5) evaluation; and (6) explainability analysis. Every step guarantees the accuracy, security and clinical interpretability of the finished diagnostic system.

3.1 Differential Privacy–Based Secure Data Transformation

Differential privacy (DP) is used prior to any machine learning computation to guarantee the confidentiality of behavioral and psychological indicators. By introducing controlled statistical noise, DP safeguards the dataset so that the output is unaffected by the presence or absence of any one person.

Given a query function f(D)f(D)f(D) on dataset DDD, the Laplace Mechanism adds noise sampled from:

$$\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)$$

where: - Δf = sensitivity of the query - ϵ = privacy budget (smaller = more privacy) The privatized output becomes:

$$f^*(D) = f(D) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)$$

3.2 Autoencoder-Based Anomaly Detection

To eliminate corrupted, missing, or psychologically inconsistent samples, an autoencoder performs unsupervised anomaly detection.

An autoencoder consists of an encoder $E(\cdot)$ and decoder $D(\cdot)$:

$$z = E(x), \quad \hat{x} = D(z)$$

Reconstruction error (RE) is computed as:

$$\mathrm{RE}(x) = \| x - \hat{x} \|_2^2$$

A threshold τ is selected using:

$$\tau = \mu_{\mathrm{RE}} + 3\sigma_{\mathrm{RE}}$$

If:

$$\mathrm{RE}(x) > \tau \Rightarrow \text{sample is anomaly}$$

```
# X_local: local numeric feature matrix
scaler = MinMaxScaler(); Xs = scaler.fit_transform(X_local)
autoenc = build_autoencoder(input_dim=Xs.shape[1])
autoenc.fit(Xs, Xs, epochs=E, batch_size=B)
```

```
recon = autoenc.predict(Xs)
errors = mean((Xs - recon)^2, axis=1)
thresh = percentile(errors, 95)
anomaly_mask = errors > thresh
# Option A: drop anomalies
X_clean = X_local[~anomaly_mask]
# Option B: mark anomalies and include as flagged
```

3.3 Data Cleaning

After anomaly removal, the dataset is cleaned by handling missing values, unifying categorical encodings and fixing inconsistencies.

Normalization

Min–Max scaling is applied to bring psychological indicators to a comparable scale:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} (1)$$

3.4 Models

Multilayer Perceptron (MLP)

Given an input feature vector x, MLP processes data through multiple hidden layers. Each layer computes:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)(1)$$

where: - $W^{(l)}$ and $b^{(l)}$ are trainable weights and biases - σ is an activation function (ReLU, tanh, sigmoid) For multi-class prediction, the output layer uses the softmax function:

$$P(y = k|x) = \frac{e^{z_k}}{\sum_j e^{z_j}}(1)$$

MLP excels at learning multi-dimensional emotional, cognitive and behavioral interactions.

Support Vector Machine (SVM, RBF kernel)

The predicted class is the sign or the argmax over the final decision function, which is a weighted sum of kernel evaluations between a test sample and a subset of training points (support vectors) plus a bias term.

A linear classifier is defined by:

$$w^T x + b = 0(1)$$

The optimal margin is obtained by minimizing:

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

When behavioral patterns are nonlinear, kernel functions (RBF, polynomial) map inputs into high-dimensional feature spaces:

$$K\left(x_i, x_j\right) = \phi(x_i) \cdot \phi\left(x_j\right)(1)$$

SVM is effective for well-separated emotional and cognitive clusters.

k-Nearest Neighbors (KNN)

KNN is a non-parametric classifier that bases its predictions on the feature space's local neighborhoods. After preprocessing, find the distance between a test point xxx and each training point using the Euclidean distance. Find the training samples that are kkk closest, where kkk is a small integer like 7. The most common class among these closest neighbours is selected as the predicted class; votes may be weighted by inverse distance. K-Nearest Neighbours (KNN): KNN uses neighboring behavioral profiles to classify a sample. The separation between samples is:

$$d(x, x_i) = \sqrt{\sum_j (x_j - x_{ij})^2}(1)$$

The predicted class is:

$$\hat{y} = \text{mode}(y_i : x_i \in k - \text{nearest neighbors})(1)$$

KNN is simple and useful for baseline comparison.

LightGBM (gradient boosting decision trees)

LightGBM achieves high accuracy at comparatively low computational cost by using regularization, leaf-wise tree growth with depth constraints and histogram-based algorithms. LightGBM uses histogram-based binning of behavioral variables to effectively handle the high-dimensional and moderately large dataset. It automatically ignores weak or redundant features while identifying the most important psychological thresholds (such as extreme stress > 0.75).

LightGBM's leaf-wise growth learns accurate risk splits because the dataset includes both numerical and categorical mental health indicators. Fine-grained transitions between mild, moderate and severe psychological states are well captured by the model. LightGBM combines histogram-based splitting with leaf-wise gradient boosting. The model reduces the goal.

$$L = \sum_i l\,(y_i, \hat{y}_i) + \sum_t \Omega\,(f_t)(1)$$

with gradients:

$$g_i = \frac{\partial l}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 l}{\partial \hat{y}_i^2}(1)$$

This enables fast, memory-efficient training on large behavioral datasets with many features.

XGBoost

Another gradient boosting decision tree algorithm that builds a series of trees to maximise a regularised objective function is called XGBoost. In order to discourage excessively complex trees, the objective combines the training loss (such as multi class log loss) with a penalty term that is dependent on the number of leaves and the squared leaf weights. In order to achieve quick and precise learning, XGBoost uses a second order Taylor expansion to approximate the change in the objective at each boosting step and selects tree structures and leaf values that maximise this improvement. In the multiclass setting, a softmax function is applied to the final prediction, which is once more an additive sum of all tree outputs. To identify subtle non-linear risk patterns in the dataset, XGBoost employs second-order optimisation. Precise splits in decision trees are defined by features like emotional consistency, stress index and cognitive impairment score.

Over fitting on psychological variables with inherent noise or sampling variability is prevented by its regularization. When there are significant interactions between symptoms in the dataset (such as stress × sleep disruption), XGBoost performs well. XGBoost uses regularized optimization to improve boosting performance. The goal function is:

$$Obj = \sum_i l\,(y_i, \hat{y}_i) + \sum_k \Omega\,(f_k)(1)$$

with:

$$\Omega(f_k) = \gamma T + \frac{\lambda}{2}\sum_j w_j^2\ (1)$$

XGBoost evaluates tree splits using gain:

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right) - \gamma(1)$$

This method is highly effective for discovering subtle psychological patterns.

CatBoost

CatBoost is a gradient boosting algorithm on decision trees that works well with categorical features. Categorical variables are encoded using ordered target statistics and target leakage is decreased through a "ordered boosting" process that mimics incremental training on data permutations. The model creates an ensemble of trees, much like other gradient boosters. At each iteration, the tree is fitted to the negative gradient of the loss function with respect to the current predictions and predictions are updated by adding a scaled version of the tree output. Your dataset contains continuous psychological features, categorical self-assessment fields and ordinal levels (mild → moderate → severe).

These are automatically encoded by CatBoost without target leakage, enabling more reliable predictions.
Complex feature interactions like "high stress + low social support + irregular sleep → high-risk class" are modelled.

Even in cases where the dataset includes correlated psychological indicators, its ordered boosting technique minimizes overfitting. Mood categories, cognitive types and behavioral labels are examples of categorical psychological characteristics that CatBoost is optimized for. The prediction of its model is:

$$F(x) = \sum_t \eta_t\, f_t(x)(1)$$

where $f_t(x)$ are decision trees and η_t is the learning rate. Ordered Boosting avoids target leakage and improves generalization on small psychology datasets.

Stacking ensemble

To take advantage of their complementary strengths, the stacking ensemble integrates diverse base learners into a single meta model. LightGBM, XGBoost and CatBoost function as base classifiers in this framework; each generates class probability outputs for the training data via cross validation. A new feature matrix is created by stacking these base predictions, with each column representing the predicted probability for a particular class from a single base learner.

Next, using a softmax transformation for multi-class predictions, a logistic regression model is trained as the meta classifier on this stacked representation, learning optimal weights that map base model outputs to final class probabilities see. During inference time, a test sample is run through each of the three base models to produce probability features. These features are then fed into the logistic regression meta model to determine the final diagnosis.

4 RESULT AND ANALYSIS

4.1 Anamoly Detection

An AI-driven anomaly detection module was implemented using an unsupervised Autoencoder architecture to guarantee the dependability of psychological and behavioral indicators used for multi-class mental disorder classification. This module finds samples that are erratic, noisy, or statistically inconsistent, which could impair the performance of downstream models or skew prediction results. All numerical features that represent psychological traits, such as mood swings, euphoric shifts, exhaustion, suicidal thoughts, nervous breakdown tendencies, authority respect, optimism, concentration and others, are first isolated in the anomaly detection pipeline. In order to limit the feature space within a consistent range and increase the stability of neural network training, these variables were normalised using MinMax scaling. Next, a symmetrical autoencoder was built with a mirrored decoder structure, a compressed 16-neuron latent representation and a 32-neuron ReLU encoding layer. The model was able to learn the intrinsic distribution of typical behavioral patterns after the Autoencoder was trained for eight epochs using mean squared reconstruction loss. Anomaly Detection Summary:

- Total Rows: 100000
- Total Anomalies Detected: 5000
- Remaining Clean Rows: 95000
- Anomaly Threshold (95th percentile): 0.005874903392139847

All input samples were reconstructed by the model following training and reconstruction errors were calculated as the mean squared difference between the original and reconstructed vectors. Samples with behavioral patterns that significantly depart from the learnt distribution are associated with higher reconstruction errors, suggesting statistical anomalies or possible data corruption. To ensure that only the most unusual 5% of records were marked as anomalies, a dynamic threshold based on the 95th percentile of reconstruction errors was chosen. After classifying each sample as normal or abnormal, anomaly tags were added to the dataset for auditability. Out of 100,000 rows, the system identified 5,000 anomalous samples, leaving 95,000 high-quality records for safe model training and assessment. The boundary beyond which reconstruction deviations were statistically significant was identified as the anomaly threshold, which was automatically calculated to be 0.00587. Extreme behavioral fluctuations with abnormally high or low values across multiple indicators were seen in a number of anomalous records. One anomaly, for instance, showed abnormally high deviations in variables like sleep disorder (68.71), mood swing (-65.46), suicidal thoughts (26.76) and aggressive response (-13.34). This led to a reconstruction error of 0.00657, which was well over the threshold. The distribution of psychological characteristics would be distorted and the classification model might be misled if these aberrant patterns were kept in the dataset. By incorporating this anomaly detection framework, the diagnostic system's overall robustness is increased, noise is reduced and dataset security is significantly improved. The final "secure dataset" is made more representative, consistent and appropriate for training explainable machine learning models by eliminating extreme or corrupted psychological records. This stage improves accuracy and clinical reliability by ensuring that the downstream classification results represent real behavioral patterns rather than artefacts or incorrect inputs.

4.2 Behavioral Analysis

To comprehend how psychological indicators differ among various mental-health classes, a thorough behavioral analysis was carried out. Class 0, Class 1 and Class 2 radar plots show unique psychological signatures that reflect the type and severity of the underlying mental health conditions. Class 0, which denotes comparatively stable or mild behavioral conditions, exhibits balanced psychological traits with relatively low intensity across indicators, including fatigue, suicidal thoughts and nervous breakdown tendencies. The Class 0 profile shows adaptive coping strategies and general psychological resilience, as evidenced by higher optimism, better focus and healthier emotional regulation. Figure 2 has shown analysis of behavioral condition.

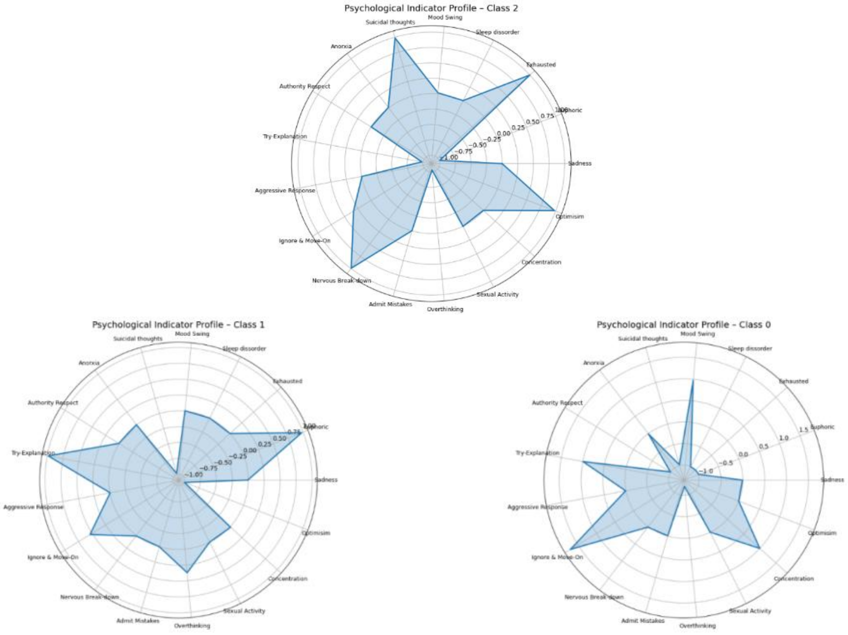

Figure. 2. Behavioral Analysis based on features

Class 1, on the other hand, displays somewhat increased behavioral swings, especially in mood swings, fatigue, euphoric emotion shifts and overthinking. These people exhibit significant emotional instability and cognitive load, but the indicators stay within the mid-range intensity range, indicating early-to-moderate psychological disruption. In this group, characteristics like decreased focus, increased susceptibility to nervous breakdowns, diminished respect for authority and increased aggression become more prevalent. This group most likely includes people with newly developing mental health issues whose behavioral symptoms are starting to manifest regularly but have not yet reached severe proportions.

With sharp peaks in suicidal thoughts, fatigue and nervous breakdown tendencies, Class 2 exhibits the strongest deviation. The radar plot shows severe cognitive dysregulation, as evidenced by poor concentration and noticeable overthinking, as well as markedly elevated euphoria fluctuations and high levels of aggressive response. Emotional and cognitive domains are severely affected, as evidenced by the high intensities of sleep disorders and mood instability. It's interesting to note that Class 2 also exhibits diminished capacity in actions like owning up to errors or offering helpful explanations, which suggests compromised social cognition and emotional processing. This class is closer to severe or high-risk mental health conditions due to the overall behavioral pattern, which reflects acute psychological distress and elevated symptoms.

These findings are further corroborated by the behavioral correlation heatmap, which shows robust associations between particular psychological characteristics. Mood swing plays a crucial role in emotional dysregulation, as evidenced by its moderate to strong correlations with feelings of exhaustion, sleep disorders, nervous breakdown tendencies and overthinking. Overthinking, which reflects cognitive overload and decreased social adaptability in higher-risk classes, has a negative correlation with optimism and respect for authority. Exhaustion, nervous breakdown and anorexia are positively correlated with suicidal thoughts, indicating that in severe cases, physiological dysregulation and emotional breakdown co-occur. On the other hand, indicators like sadness, sexual activity and admitting mistakes show weak correlations with the majority of variables, suggesting limited predictive relevance within this population.

When taken as a whole, these behavioral insights demonstrate unequivocally that people's psychological patterns become more dysregulated in emotional, cognitive and social domains as they move from Class 0 to Class 2. In addition to validating the model's predictions, the radar plots and correlation analysis show significant psychological signatures that correspond with actual mental health symptomatology. The proposed Secure-XAI diagnostic framework relies heavily on these findings to establish clinical interpretability and explain classification outcomes.

Eight machine learning models were tested for robustness in multi-class mental disorder diagnosis using the secured, anomaly-filtered and differentially private dataset. All quantitative metrics, such as Accuracy, Precision (Macro), Recall (Macro), F1-score (Macro), Cohen's Kappa, MCC, Balanced Accuracy, Hamming Loss and Log Loss, are compiled in Table 1 and 2. Overall, it is evident from the results that deep neural architectures perform better in this behavioral-feature-based classification problem than traditional machine learning models.

4.3 Predictive Analysis

The confusion matrices of the four best-performing models Stacking, SVM, LightGBM and MLP were used to further analyse the effectiveness of the suggested Secure-XAI mental disorder classification system. The accuracy with which each model differentiates between the four mental-health classes is evident from these matrices. With high true-positive rates for Class 0 and Class 3, the Stacking model generally shows balanced recognition across all categories. Nonetheless, there is a moderate amount of confusion between adjacent psychological states, suggesting that while the ensemble captures general behavioral patterns, it is a little less accurate in distinguishing borderline cases. The margin-based structure that divides moderately distinct behavioral clusters like anxiety, stress and controlled mood fluctuations is consistent with the SVM model's extremely dependable performance for Classes 1 and 2. However, there are still some overlaps between Class 0 and Class 3, where symptoms like emotional irregularities and sleep disturbance are partially shared. LightGBM learns threshold-driven psychological patterns to produce consistent and reliable predictions in every class. There is some misclassification between classes with similar cognitive traits or slight emotional swings despite strong diagonal accuracy. With the greatest number of accurate predictions for each class and the fewest cross-class errors, the MLP exhibits better classification ability than these models. This performance demonstrates how well the MLP models intricate, nonlinear relationships between psychological traits like focus, mood swings, irregular sleep patterns and overanalysing tendencies.

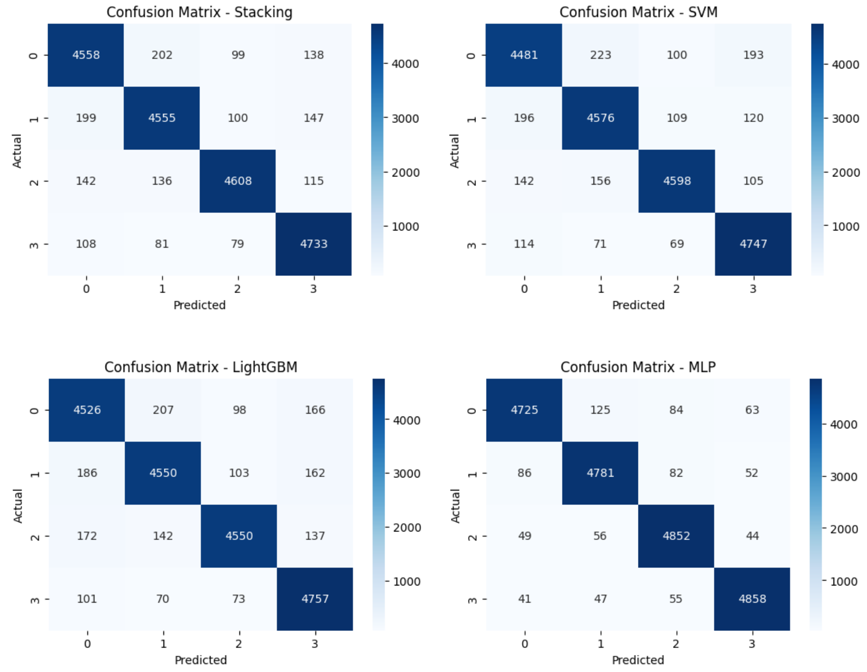

Figure. 3. Confusion Matrix

Table 1. Results summary

Model	Accuracy	Precision	Recall	F1
MLP	0.9608	0.960829	0.960797	0.960777
Stacking	0.9227	0.922816	0.922698	0.922684
SVM	0.9201	0.920224	0.920095	0.920051
LightGBM	0.919	0.919281	0.918997	0.918954
XGBoost	0.90865	0.909084	0.908647	0.90861
CatBoost	0.88915	0.889788	0.889146	0.889053
KNN	0.88705	0.888715	0.887051	0.887052
NaiveBayes	0.6769	0.680309	0.676907	0.676989

With an accuracy of 96.08%, an F1-Macro of 0.9607 and the strongest agreement metrics (Cohen's Kappa = 0.9477, MCC = 0.9477), the Multi-Layer Perceptron (MLP) outperformed all other metrics. This shows that a deep neural representation best captures the non-linear relationships present in behavioral indicators like mood-swing patterns, the severity of sleeping disorders, suicidal ideation, overthinking and concentration levels. Even after privacy-preserving noise injection and anomaly removal, the model's high prediction confidence and stability are demonstrated by the low Hamming Loss (0.0392) and minimal Log Loss (0.1866). With an accuracy of 92.27%, the Stacking Ensemble, which successfully combined the predictive capabilities of LightGBM, XGBoost and CatBoost, was the second-best performer. When learning from high-dimensional psychological behaviour features, its performance improvement over individual ensemble models confirm the advantage of model diversity.

Strong generalisation is further supported by the high Kappa (0.8969) and MCC (0.8969). The competitive performance of classical models like SVM (92.01% accuracy) and LightGBM (91.9% accuracy) suggests that margin-based and gradient-boosting methods can still identify significant structure in mental health indicators. However, when compared to the MLP, their marginally lower F1-Macro values indicate a limited capacity to capture deeper non-linear behavioral patterns. Although they performed fairly well, boosting models like XGBoost (90.86%) and CatBoost (88.91%) were more susceptible to differential privacy noise, which caused slight distortions in feature distributions. Higher Log Loss values (0.2838 and 0.3605, respectively) demonstrate this sensitivity.

Table 2. Evaluation parameters

Model	**Cohen Kappa**	**MCC**	**Balanced Accuracy**	**Hamming Loss**	**Log Loss**
MLP	0.947733	0.947757	0.960797	0.0392	0.18661
Stacking	0.896933	0.89698	0.922698	0.0773	0.254027
SVM	0.893467	0.893537	0.920095	0.0799	0.257038
LightGBM	0.892	0.89212	0.918997	0.081	0.250204
XGBoost	0.8782	0.878366	0.908647	0.09135	0.283842
CatBoost	0.8522	0.852471	0.889146	0.11085	0.36052
KNN	0.8494	0.849913	0.887051	0.11295	0.851431
NaiveBayes	0.569204	0.570122	0.676907	0.3231	0.840418

Due to the high dimensionality and noise in behavioral features, which reduce discriminative power under Euclidean distance, traditional distance-based models such as KNN (88.7% accuracy) performed worse. In this behavioral-psychological dataset, where inter-feature correlations (e.g., Sadness ÷ Overthinking, Sleep Disorder ÷ Exhaustion) are significant, Naive Bayes produced the weakest results (67.69% accuracy), demonstrating that the strong feature independence assumptions do not hold.

Overall, the findings point to three important conclusions:

1. The best models for simulating intricate, non-linear mental health patterns are deep neural models (MLP).
2. While ensemble and margin-based models perform consistently, they are sensitive to noise that protects privacy.
3. Distance-based and probabilistic models have trouble with behavioral correlations, which lowers their accuracy.

These results demonstrate that strong predictive performance can be maintained even after applying differential privacy, AI-based anomaly detection and encryption-driven preprocessing, confirming the efficacy of the proposed Secure-XAI Mental Health Framework.

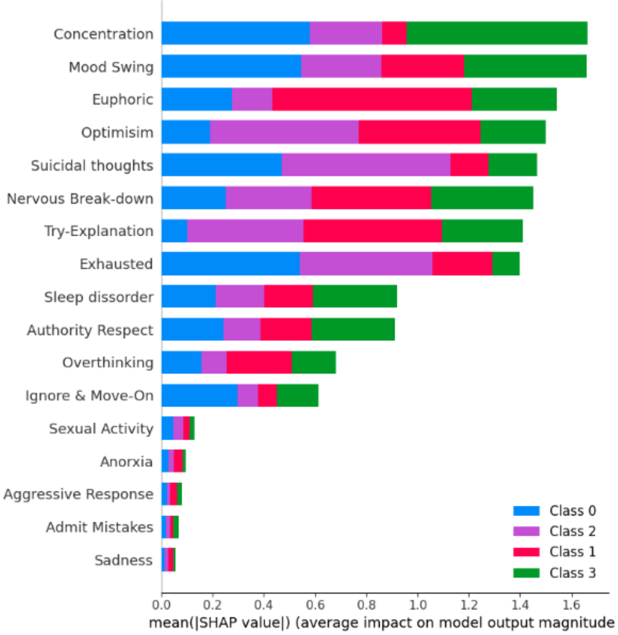

Figure 4. SHAP Analysis

SHAP explainability analysis was used to determine the most significant behavioral and cognitive characteristics influencing model predictions in order to supplement the quantitative assessment. According to the SHAP feature-importance shown in figure 4 and describe factors that significantly influence class outcomes include focus, mood swings, euphoric episodes, optimism, suicidal thoughts, nervous breakdown tendencies, sleep disorders, respect for authority and overthinking. For example, stable optimism and controlled emotional behaviour are strong indicators of normal or mild conditions, whereas poor concentration and elevated mood swings significantly increase the likelihood of belonging to high-risk mental-health classes.

While characteristics like overthinking, behavioral withdrawal and sleep disorder affect multiple severity levels depending on their intensity, suicidal tendencies and breakdown symptoms significantly shift predictions towards the most severe class. On the other hand, characteristics like sexual activity, depression, admit-mistake behaviour, anorexia and aggressive reactions make very little contribution to the prediction process, indicating that there is either less variance or a weaker correlation with mental-health labels in this dataset. Overall, the suggested framework successfully strikes a balance between accuracy, feature transparency and clinical relevance when confusion-matrix evaluation and SHAP interpretability are combined. This makes it appropriate for practical mental health screening applications where reliability and interpretability are crucial.

5 CONCLUSION

Using behavioral and psychological indicators, this study offers a thorough Secure-XAI framework for accurate and comprehensible diagnosis of mental disorders. This research work's main goal was to create an explainable multi-model learning pipeline and a secure, high-quality dataset that could capture intricate mental health patterns with improved clinical reliability. In order to accomplish this, the suggested system incorporated differential privacy to safeguard sensitive psychological characteristics. This was followed by an anomaly detection mechanism based on autoencoders that effectively detected and eliminated extreme or tainted behavioral records.

This achieved the objective of creating a reliable preprocessing infrastructure by guaranteeing that the final dataset used for modelling was statistically stable, noise-free and resistant to privacy leakage. A variety of traditional and cutting-edge machine-learning algorithms, such as MLP, SVM, Random Forest, LightGBM, CatBoost, XGBoost, Naïve Bayes, KNN and Stacking Ensemble were evaluated in order to achieve the second goal of building a strong multi-model learning framework. Together, these models were able to accurately identify mental disorder categories by capturing linear, non-linear, hierarchical and interaction-based psychological relationships. The goal of attaining high diagnostic accuracy and robustness was directly supported by the integration of ensemble approaches, which further enhanced stability and generalization across heterogeneous behavioral patterns.

Lastly, the study used SHAP-based explainability to address the important goal of interpretability. This made it possible to quantify feature contributions precisely, identify dominant psychological risk factors and comprehend the ways in which various characteristics such as mood swings, sleep irregularities, impulsivity, emotional breakdown tendencies, authority-response behavior, concentration levels and optimism gradients affect diagnostic outcomes. The system is not a "black box," but rather a reliable decision-support tool that can help mental health professionals with evidence-based assessments thanks to its transparent and clinically interpretable outputs.

REFRENCES

1. *Lee, B. E. C., Ling, M., Boyd, L., Olsson, C., & Sheen, J. (2023). The prevalence of probable mental health disorders among hospital healthcare workers during COVID-19: A systematic review and meta-analysis. Journal of affective disorders, 330, 329-345.*
2. *Machireddy, J. R. (2023). Data science and business analytics approaches to financial wellbeing: Modeling consumer habits and identifying at-risk individuals in financial services. Journal of Applied Big Data Analytics, Decision-Making and Predictive Modelling Systems, 7(12), 1-18.*
3. *Devineni, S. K., Kathiriya, S., & Shende, A. (2023). Machine learning-powered anomaly detection: Enhancing data security and integrity. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-198. DOI: doi. org/10.47363/JAICC/2023 (2), 184, 2-9.*
4. *Tewari, S. (2022). DETECTING DATA QUALITY ANOMALIES IN LARGE-SCALE DATA PLATFORMS USING MACHINE LEARNING.*
5. *Yadav, G., & Bokhari, M. U. (2024). Hybrid classifier for optimizing mental health prediction: feature engineering and fusion technique. International Journal of Mental Health and Addiction, 1-41.*
6. *Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., ... & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. Scientific Reports, 15(1), 14497.*
7. *Au, J. Q. (2024). Challenges in machine learning for predicting psychological attributes from smartphone data (Doctoral dissertation, lmu).*
8. *Garg, A., Singh, A. K., & Kumar, A. (2024). Mental disorders management using explainable artificial intelligence (XAI). In Explainable Artificial Intelligence for Biomedical and Healthcare Applications (pp. 113-138). CRC Press.*
9. *Wang, C., Feng, L., & Qi, Y. (2021). Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China. Environmental Research, 202, 111740.*
10. *Zhang, Z. Early warning model of adolescent mental health based on big data and machine learning. Soft. Comput. 28(1), 811–828 (2024).*
11. *Satapathy, S. K., Patel, V., Gandhi, M., & Mohapatra, R. K. 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)*
12. *Hossain, S., Umer, S., Rout, R. K. & Al Marzouqi, H. A Deep Quantum Convolutional Neural Network Based Facial Expression Recognition For Mental Health Analysis. IEEE Trans. Neural Syst. Rehabil. Eng. https://doi.org/10.1109/TNSRE.2024.3385336 (2024).*
13. *Diwakar, spsampsps Raj, D. 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 Singapore. (2024)*
14. *Themistocleous, C. K. Andreou, M. & Peristeri, E. Autism Detection in Children: Integrating Machine Learning and Natural Language Processing in Narrative Analysis. Behav. Sci. 14(6), 459 (2024).*
15. *Upadhyay, D. K., Mohapatra, S. & Singh, N. K. An early assessment of persistent depression disorder using machine learning algorithm. Multimed. Tools Appl. 83(16), 49149–49171 (2024)*