Intelligent Depression Classification Based on Hybrid Models Using Actigraphy

Thamina anzum A[1], Thirumahal R[2], Gobika R[3], Vinithaa P[4], Naveen Ragav K[5], Naveen P[6]
[1]*Corresponding authors Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India*
[3,4,5,6]*Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India*
[2]*Assistant Professor (Sl Gr), Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India*

***Abstract*—Depression is a serious mental health problem that affects an individual's emotion and daily activities which requires early and accurate diagnosis. Traditional methods depend on clinical interviews and questionnaires, which may not capture continuous behavioural changes. By using wearable devices, actigraphy-based data provides continuous values of human behaviour that is used for automated depression assessment. In this work, we develop an "Intelligent depression classification system" using actigraphy data based on two hybrid machine learning models: CatBoost-Artificial Neural Network (CatBoost-NN) and LightGBM-Artificial Neural Network (LightGBM-NN). The system classifies individuals into bipolar I, bipolar II and unipolar categories. Tree based models (CatBoost and LightGBM) are used to learn from clinical patient information, while Neural Network is used for identifying deep non-linear relationships from activity-based and behavioural features. The dataset contains minute-level data collected from depressed and healthy subjects, collected using wearable actigraphy sensors. The hybrid models combine the strengths of both approaches to improve classification. Explainable AI techniques using SHAP are used to interpret model predictions and identify key features that result in depression detection. The proposed hybrid approach aims to achieve higher accuracy when compared to traditional singlemodel methods, providing the effectiveness of combining wearable sensor data with hybrid and explainable machine learning techniques for automated depression classification.**

***Index Terms*—Depression classification; Actigraphy data; Wearable sensors; Hybrid machine learning; Explainable AI; Mental health diagnosis**

I. INTRODUCTION

A person's emotions, behaviour, and day-to-day activities are all affected by depression, a severe mental illness. It is the most common cause of disability and needs to be checked out early in order to be effectively treated. Standard diagnostic methods depend on questionnaires and clinical interviews, which are restricted to quick observation times. Traditional ways of identifying depression are difficult in the real world. The depression depends on the patient itself and it is not useful while classifying the type of depression. Physicists check the patient only for few minutes and not regularly, so many important reasons for depression are missed for weeks and months.

Actigraphy-based monitoring can be used as a solution for mental health evaluation because of developments in wearable technology. Actigraphy devices continuously record activity data, which provides valuable information such as sleep patterns, physical activity levels, and abnormalities of the circadian rhythm, all of which are associated with depressive disorders. It automatically identifies depression and helps in clinical decision making by employing machine learning techniques with these time-series signals. Machine learning is very useful for mental health issues. There are old techniques which cannot identify different type of depression and also don't use actigraphy data. So, these systems can't explain why the patient is depressed and the reason of depression also not identified and cannot be used in real-world.
To address these limitations, this work proposes an Intelligent Depression Classification System using Actigraphy Data based on two hybrid machine

learning models: CatBoost with Artificial Neural Network (CatBoost-NN) and LightGBM with Artificial Neural Network (LightGBM-NN). 1

The proposed system classifies individuals into Bipolar I, Bipolar II and Unipolar Depression with improved accuracy. The hybrid learning approach combines the strengths of tree-based models in handling structured patient data and the capability of neural networks in learning deep non-linear behavioural patterns from activity-based features.

The motivation behind this work lies in overcoming the limitations of traditional diagnostic methods and improving the reliability of automated depression classification. Continuous actigraphy data enables long-term behavioural monitoring, which provides more meaningful insights compared to short clinical observations. In addition, the use of Explainable Artificial Intelligence (XAI) techniques such as SHAP improves model transparency by identifying the most influential features contributing to predictions, thereby increasing trust among clinicians.

The main contributions of this paper are as follows:

- Development of a hybrid CatBoost-NN and LightGBM-NN framework for multi-class depression subtype classification
- Integration of actigraphy-based behavioural features with structured clinical data
- SHAP-based explainability for improved model interpretability and clinician trust
- Comparative performance evaluation of both hybrid models using multiple clinical metrics

This study develops and evaluates a multivariable machine learning-based clinical prediction model for automated classification of depression subtypes using actigraphy and clinical features.

II. LITERATURE REVIEW

[1] Patil et al. presented a hybrid approach in classifying depression using a Random Forest classifier and an ANN model on activity signals collected from wearable devices. The Random Forest classifier handles the non-linear relationships in activity features, and the ANN model is based on activity patterns related to depressive behaviours. The results showed a better performance in terms of accuracy by using a hybrid approach when compared to individual models. However, the study was limited by the small dataset size.

[2] Dong at al. developed a machine learning model that does depression risk prediction based on demographic, lifestyle, dietary, and familial risk factors using the NHANES dataset. The results showed 11 different machine learning models were compared in this research, and Random Forest showed better performance compared to other machine learning models. The results were compared using SHAP values. However, the dataset is limited to a Korean population.

[3] Choi et al. presented a machine learning approach in classifying depression level using actigraphy-based circadian rhythm features collected from a wearable device. Multiple MACHINE LEARNING models were used in this research, and XGBoost among them showed better performance in classifying depression level compared to other models. The results showed two days of actigraphy data are sufficient in classifying depression level. However, the results were limited to a Korean population and did not use any deep learning models.

[4] Huang et al. presented a machine learning framework in classifying postpartum depression risk using psychological factors based on an explainable machine learning approach. XGBoost showed better performance in classifying postpartum depression risk compared to other machine learning models by selecting features using LASSO and Boruta methods. However, the dataset is limited as data were collected from a single hospital.

[5] Aleem et al. presented a review of various machine learning and deep learning models for diagnosing depression. This paper mainly identifies the features that majorly contribute to depression and also highlights the efficiency of deep learning models. Though this paper provides valuable insights, it doesn't actually implement any models and no standard dataset was used.

[6] Shah et al. developed a methodology for detecting depression by analysing social media text using refined Large Language Models (LLMs) like GPT-3.5 and LLaMA-2. The work shows better performance than traditional NLP methods by identifying depressionrelated verbal patterns using transformer-based models. Yet the model lacks explainability, is restricted to text-based binary classification, and does

not provide any information about the type of depression.

[7] Kerasiotis et al. introduced a transformerbased method to detect depression from social media posts. Their approach uses text understanding along with extra language and user-related features to improve the analysis. Models like BERT are good at understanding the meaning and context of text, which helps improve classification results. However, this approach only uses text data and classifies results into just two categories. It also does not include explainable AI methods or real behavioural data, which makes it less useful in real clinical settings.

[8] Khan et al. introduced a machine learning-based depression model using EEG signals from the MODMA dataset. This study used classifier modes like Best First Tree, KNN, and AdaBoost to extract EEG features to obtain higher accuracy. This work shows that physiological signals can help detect depression. However, it depends on special EEG equipment, which makes it hard to use on a large scale. It also lacks explainability, does not use hybrid models, and only focuses on limited types of depression analysis.

[9] Jamali et al. applied machine learning techniques to large scale health survey data from the NHANES dataset to identify depression predictors. Demographic, lifestyle, and health related features were analysed to build classification models and determine risk factors associated with depression. Although it works well for large groups, the approach relies on self-reported data and does not support continuous monitoring.

[10] Teferra presented a review of text-based depression prediction methods that uses NLP techniques on social media data. This paper discusses traditional machine learning and transformer-based approaches and highlights challenges such as data imbalance, and lack of explainability. This study is limited to text-based analysis and uses traditional machine learning models and does not speak about hybrid models.

[11] Patel et al. explored in detecting depression moods using daily step count that are collected from wearable devices. The study used explainable machine learning methods to find important behaviour patterns, like low activity levels and disturbed sleep cycles. Models like CatBoost, XGBoost, and random forest are trained on actigraphy data that are collected from wearables. However, the study was limited to binary classification and does not help identify depression subtypes such as Bipolar I, Bipolar II, and Unipolar depression.

[12] Ramirez et al. presented a model for detecting depression disorder using minute level actigraphy data collected from wearable devices. Both traditional machine learning models and deep learning models based on convolutional neural networks were evaluated, where CNN performed efficiently by learning patterns when compared to other models. This study shows how deep learning works with wearable datasets, but supports only binary classification and also it lacks explainability.

[13] Zhang et al. used motor activity signals from actigraphy sensors to develop a machine learning-based model for classifying depression. Support Vector Machines, Random Forest, and K-Nearest Neighbours are the machine learning models that were used for extracting the temporal coefficients that were extracted from autoregressive time-series modelling. In this paper, feature engineering is done manually and also it doesn't explain about how deep learning models behave with this data.

[14] Park et al. explored transfer learning methods to overcome the problem of limited data in actigraphy based data. They fine-tuned the pretrained neural network models on the new dataset so that it can perform well even with the limited data. However, these models are like black boxes and they don't provide any explanation for their prediction, hence this approach can't be trusted for clinical use. Also, this model does not use hybrid methods and does not perform multi class classification of depression.

[15] Lopez et al. conducted a review of Al-based systems using wearable devices for depression detection. In addition to machine learning and deep learning techniques like Random Forest, SVM, CNN, RNN, and gradient boosting, the study includes actigraphy and physiological data. The authors highlighted the need for advanced and explainable models by pointing out important gaps, such as the inadequate explainability of deep learning models, the absence of hybrid techniques, and the lack of multi-class and subtype categorization.

[16] Kim et al. introduced a depression detection model that combines the features of both LLM(Large Language Model) and traditional machine learning models trained on social media text. This approach

focuses on explainability of identifying meaningful linguistic and psychological features which contribute to prediction. This approach provides high performance in both binary depression classification and severity level classification. However, using fixed LLM features may limit its ability to adapt to changing language patterns.

[17] Chen et al. compared various traditional machine learning modes, transformer-based models and large language models for detecting depression and anxiety from text-based data. The results showed that LLMs gave higher performance than other models as it could understand context and meaning efficiently. The study also points out a trade-off between performance, cost, and explainability, and suggests the need for hybrid models.

[18] N.Zong et al. proposed a hybrid SBERT-CNN model for detecting depression in Reddit users. SBERT (Sentence BERT) helped in finding the meaning of the test and other model CNN - helped in capturing the important pattern from the posts. The final prediction is made by combining the results of both the models by observing various posts of the user. This approach lacks explainability and lacks symptom level or severity-based classification.

[19] Farruque et al. presented a semi - supervised model for detecting depression from social media text. This model uses both labelled and unlabelled data; this makes it more suitable for solving problems of limited data and thereby increases its performance. However, this model is more complex and it gives only yes/no results rather than saying the type of depression and hence this can't be trusted for real time use.

[20] Baydili et al. proposed a deep learningbased model with CWINCA feature selection to detect depression and suicidal behaviour from social media data. CWINCA stands for Chaotic Whale Optimization Algorithm with Neighbourhood Component Analysis. This CWINCA features selection technique helps in selecting most important data and removing unnecessary data and thereby increase the accuracy. Language model features are combined with SVM model to achieve high performance with limited data and reduces overfitting. While the model performs well, it mainly provides statistical explanations rather than meaningful clinical insights.

From the reviewed literature, it is evident that existing studies mainly focus on binary depression detection, text-based analysis, or single-model approaches. Very few studies address multi-class depression subtype classification using continuous actigraphybased behavioural data. Moreover, limited attention has been given to hybrid machine learning frameworks that combine treebased models with deep learning architectures while ensuring model interpretability. These research gaps motivate the proposed hybrid CatBoost-NN and LightGBM-NN framework with SHAP-based explainability.

III. SYSTEM ARCHITECTURE

3.1 Training Setup

The proposed system uses a hybrid learning framework for depression type classification by combining tree-based models with a shared neural network feature extractor. Two hybrid architectures were implemented and compared:

- CatBoost + Neural Network (NN)
- LightGBM + Neural Network (NN)

The primary objective of this setup is to combine the strengths of both approaches.

The tree-based models are highly effective in learning from structured clinical and categorical patient information, while the shared neural network is used to capture complex non-linear relationships from numeric severity-based features such as episode duration and MADRS scores.

Following window-based feature extraction from the 55 actigraphy files, a total of 11,250 samples were generated. These samples were distributed across three depression subtypes based on the afftype label from scores.csv.

The three depression categories considered in this study are:

- Bipolar I
- Bipolar II
- Unipolar Depression

This balanced setup ensures fair learning across all classes and prevents bias toward any single depression type.

Before training, the dataset was divided into:

- clinical categorical features
- numerical severity features
- target diagnosis labels

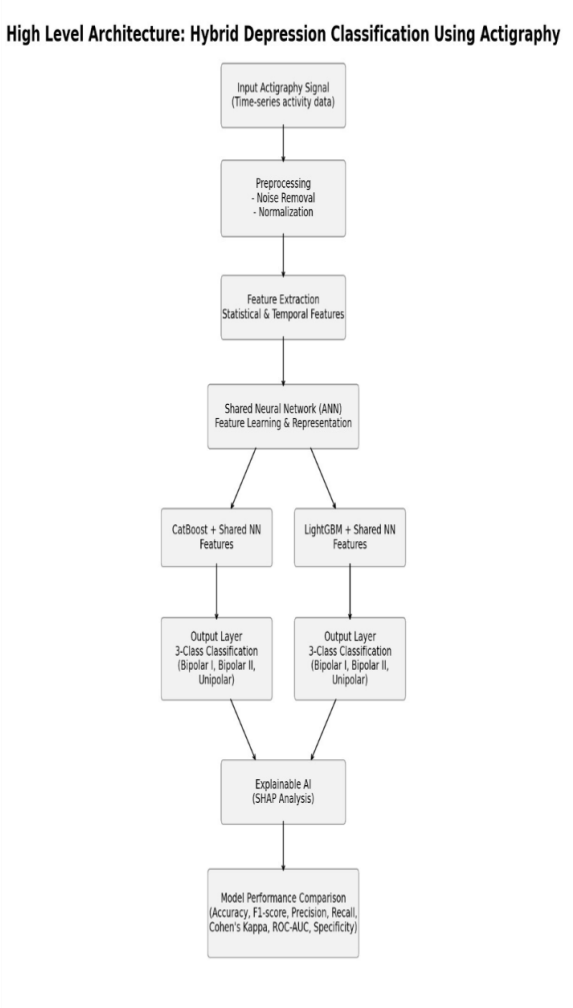

Fig. 1 System architecture of the proposed hybrid depression classification framework

Fig. 1 shows the overall architecture of the hybrid models. The numeric features were normalized using StandardScaler, ensuring zero mean and unit variance for stable neural network learning. A shared multi-layer perceptron (MLP) was used for numeric feature learning.

The network consists of:

- Input layer
- Hidden Layer 1 – 32 neurons
- Hidden Layer 2 – 16 neurons
- ReLU activation functions

The output from the final hidden layer is used as a latent feature representation, which is then combined with probability outputs from CatBoost and LightGBM.

This shared feature representation ensures:

- fair comparison
- reduced computational overhead • better generalization

Both hybrid models were trained on the same train-test split, ensuring a controlled and unbiased evaluation.

3.2 Evaluation Metrics

When evaluating models for a multi-class medical dataset, several performance metrics were considered together rather than relying only on accuracy. Since the objective is to correctly classify Bipolar I, Bipolar II, and Unipolar depression, using multiple metrics gives a clearer understanding of model performance. Below are the metrics computed and the reason for using them:

- Accuracy: the fraction of all patient records that are correctly classified. Accuracy gives the overall performance of the model and is easy to understand. However, relying only on accuracy may sometimes hide misclassification between clinically similar depression classes.

- Precision (per-class): of all the samples predicted as a particular depression type, how many actually belong to that class. Precision helps in understanding how trustworthy the model predictions are, especially when false classification between depression subtypes may lead to incorrect treatment decisions.

- Recall (per-class): of all the actual cases belonging to a specific depression class, how many were correctly identified by the model. In medical diagnosis, recall is highly important because missing an actual depression subtype can directly affect clinical treatment and patient care.

- F1-score (per-class): combines precision and recall into a single balanced metric. It is particularly useful when both false positives and false negatives need to be minimized.

- ROC-AUC (one-vs-rest): measures the ability of the model to distinguish one depression class from the remaining classes across different thresholds. This metric is useful for evaluating the discriminative strength of the hybrid models.

• Cohen's Kappa: measures the agreement between predicted and actual depression classes while accounting for agreement occurring by chance. This is particularly useful in healthcare prediction tasks where reliability and consistency are important.

• Specificity: measures how well the model correctly identifies negative cases for each class. This is useful in ensuring that the model does not incorrectly classify patients into the wrong depression subtype.

• Confusion Matrix: used to visually analyse correct and incorrect predictions for each depression category. It helps identify which depression classes are more likely to be confused with one another, such as Bipolar II and Unipolar depression.

In short, accuracy is used as an overall performance measure, while recall, F1-score, Cohen's Kappa, and ROC-AUC are considered as the primary evaluation metrics for selecting the best hybrid model. Confusion matrices are additionally used to validate the model predictions visually.

IV. IMPLEMENTATION

4.1 Dataset Description

The Depresjon Dataset, originally published by the Simula Research Laboratory and publicly available on Kaggle, is used in this study. This dataset was specifically designed to support automatic detection and classification of depression states based on continuous wearable sensor data.

The dataset is organized into two participant groups and one clinical score file. The condition folder contains actigraphy recordings from 23 patients diagnosed with depressive disorders, the control folder contains recordings from 32 healthy subjects, and scores.csv provides the corresponding clinical and demographic information for each participant, yielding 55 actigraphy files in total.

For each participant, a dedicated CSV file records time-series motor activity captured by an actigraph wrist-worn sensor at oneminute intervals. Each file contains three attributes: a timestamp indicating the minute of measurement, the corresponding date, and the raw activity count representing physical movement intensity recorded by the device.

Table 1. Clinical and demographic features from Scores.csv

Feature	Description
days	Number of days of actigraphy measurement
gender	Biological sex (female / male)
age	Age group of the participant
afftype	Depression subtype: 1 = Bipolar II, 2 = Unipolar, 3 = Bipolar I
melanch	Presence of melancholic features (yes / no)
inpatient	Treatment setting (inpatient / outpatient)
edu	Education level grouped in years
marriage	Marital status
work	Employment status
madrs1	Initial MADRS score at time of admission
madrs2	Final MADRS score at follow-up

The scores.csv file supplements the actigraphy data with structured clinical and demographic attributes. Table 1 summarizes the features used in this study.

The Montgomery–Åsberg Depression Rating Scale (MADRS) is a clinically validated instrument widely used to quantify depression severity, with higher scores indicating greater symptom burden.

To prepare the dataset for machine learning, statistical and temporal features were extracted from each participant's minutelevel actigraphy time series. Features including mean activity, standard deviation, minimum, maximum, and circadian rhythm indicators were computed per participant per measurement window. These extracted behavioural features were then merged with the clinical attributes from scores.csv to form a unified feature matrix used for model training and evaluation. The target variable is the depression subtype label derived from the afftype field, encompassing three clinically distinct classes: Bipolar I, Bipolar II, and Unipolar Depression.

The dataset is split into training, validation, and testing subsets following a 70/15/15 ratio, with class proportions preserved across all splits to ensure unbiased evaluation. Given the small participant count of 55 individuals, window-based feature extraction was employed to generate a sufficient number of

samples for model training, resulting in an expanded feature matrix suitable for the proposed hybrid classification framework.

The dataset was approximately balanced across all three classes; therefore, no additional class balancing techniques were required.

4.2 Preprocessing

The proposed depression classification system was implemented using an actigraphy-based dataset containing minute-level activity recordings collected from wearable sensors. The dataset includes activity recordings from both clinically diagnosed depressed patients and healthy control subjects, along with associated clinical labels., along with associated clinical labels. The actigraphy files were processed to extract meaningful behavioral and activity-based features relevant to depression analysis.

Missing values in the clinical and actigraphy-derived features were checked prior to model development. Records with incomplete target labels were excluded, and numerical missing values, if any, were handled using appropriate preprocessing methods before normalization.

A total of 55 actigraphy files were considered for the study. Using a window-based feature extraction approach, the raw timeseries signals were segmented into smaller fixed-length windows to capture temporal behavioral patterns. From this process, a total of 11,250 samples were generated.

Each sample was assigned a class label based on the afftype field from the scores.csv file, which corresponds to the depression subtype. The three target classes used for classification are:

- Bipolar I
- Bipolar II
- Unipolar Depression

The extracted features were further divided into:

- clinical categorical features
- numeric severity-based features
- target diagnosis labels

The numerical features were normalized using StandardScaler, which transforms the data to have zero mean and unit variance. This normalization improves the stability and performance of the neural network during training.

4.3 Hybrid Model Training for Depression Subtype Classification

To improve the classification performance, a hybrid learning framework was designed by combining tree-based machine learning models with a shared neural network feature extractor.

Two hybrid architectures were implemented and compared:

- CatBoost + Neural Network (CatBoost-NN)
- LightGBM + Neural Network (LightGBM-NN)

The purpose of this hybrid setup is to combine the strengths of both approaches. Tree-based models such as CatBoost and LightGBM are highly effective in learning structured clinical and categorical data, while neural networks are capable of learning complex non-linear relationships from numeric severity-based features. Model hyperparameters for CatBoost, LightGBM, and the neural network were selected empirically based on validation performance.

A shared multi-layer perceptron (MLP) was used for numerical feature learning. The neural network consists of:

- Input Layer
- Hidden Layer 1 – 32 neurons
- Hidden Layer 2 – 16 neurons
- ReLU activation functions

The output from the final hidden layer is treated as a latent feature representation. This learned representation is then combined with the probability outputs of CatBoost and LightGBM models to generate the final prediction.

This shared representation ensures:

- fair comparison between models
- reduced computational overhead
- better generalization

Both hybrid models were trained using the same train-test split, ensuring unbiased comparison and evaluation. The final trained models were subsequently evaluated using multiple classification metrics to assess their effectiveness in distinguishing depression subtypes.

4.4 Use of AI Tools

During manuscript preparation, ChatGPT (OpenAI) was used solely for language refinement, grammar correction, readability improvement, and editorial structuring of the human-written manuscript. No AI tools were used for data preprocessing, feature extraction, model development, classification, result

generation, statistical analysis, or interpretation of findings. All scientific content, experimental design, implementation, evaluation, and conclusions are entirely the work of the authors, who take full responsibility for the final manuscript.

V. RESULTS AND DISCUSSION

The hybrid depression classification system was trained using a shared neural network for numeric clinical features, followed by two hybrid classification branches: CatBoost + Shared Neural Network and LightGBM + Shared Neural Network.

5.1 Experimental Results And Performance Analysis

The training process first learns deep non-linear relationships from numeric features such as episode duration, MADRS score 1, and MADRS score 2 using the shared neural network. The learned feature representations are then combined with the outputs of the tree-based models for final depression type classification.

```
→  data 4 git:(main) ✗ python3 train_scores_balanced.py
Loading: scores_balanced_prime_no_control_11250.csv

[Branch 2] Training shared NN on numeric features (days, madrs1, madrs2)...

[Hybrid 1] Training CatBoost branch + fusion head (CatBoost + NN)...
  CatBoost+NN test accuracy: 0.9538

[Hybrid 2] Training LightGBM branch + fusion head (LightGBM + NN)...
  LightGBM+NN test accuracy: 0.9400

============================================================
  CatBoost + Shared NN — Full Performance Metrics
============================================================

  Accuracy:          0.9538
  Cohen's Kappa:     0.9307
  ROC-AUC (OvR):     0.9970

  Classification Report:
              precision    recall  f1-score   support

  Bipolar II       0.96      0.90      0.93       750
    Unipolar       0.90      0.96      0.93       750
   Bipolar I       1.00      1.00      1.00       750

    accuracy                           0.95      2250
   macro avg       0.96      0.95      0.95      2250
weighted avg       0.96      0.95      0.95      2250

  Specificity (per class):
      Bipolar II: 0.9820
        Unipolar: 0.9487
       Bipolar I: 1.0000

  Confusion Matrix:
[[673  77   0]
 [ 27 723   0]
 [  0   0 750]]
```

Fig. 2 Classification performance metrics of the CatBoost + Shared NN hybrid model

Fig. 2 presents the training output and performance metrics of the CatBoost + Shared NN model.

The model achieved:

- Accuracy = 95.38%
- Cohen's Kappa = 0.9307
- ROC-AUC = 0.9970

The classification report shows excellent performance across all three depression classes.

- Bipolar II: Precision = 0.96, Recall = 0.90, F1score = 0.93
- Unipolar: Precision = 0.90, Recall = 0.96, F1score = 0.93
- Bipolar I: Precision = 1.00, Recall = 1.00, F1score = 1.00

The confusion matrix further confirms that Bipolar I cases are perfectly classified, while only minor confusion exists between Bipolar II and Unipolar depression.

```
============================================================
  LightGBM + Shared NN — Full Performance Metrics
============================================================

  Accuracy:          0.9400
  Cohen's Kappa:     0.9100
  ROC-AUC (OvR):     0.9956

  Classification Report:
              precision    recall  f1-score   support

  Bipolar II       0.91      0.91      0.91       750
    Unipolar       0.91      0.91      0.91       750
   Bipolar I       1.00      1.00      1.00       750

    accuracy                           0.94      2250
   macro avg       0.94      0.94      0.94      2250
weighted avg       0.94      0.94      0.94      2250

  Specificity (per class):
      Bipolar II: 0.9567
        Unipolar: 0.9533
       Bipolar I: 1.0000

  Confusion Matrix:
[[680  70   0]
 [ 65 685   0]
 [  0   0 750]]


============================================================
  SUMMARY COMPARISON
============================================================
  CatBoost + NN  — Accuracy: 0.9538  |  Kappa: 0.9307
  LightGBM + NN  — Accuracy: 0.9400  |  Kappa: 0.9100
============================================================
```

Fig. 3 Classification performance metrics of the LightGBM + Shared NN hybrid model

Fig. 3 shows the output performance of the LightGBM + Shared NN model.

The model achieved:

- Accuracy = 94.00%
- Cohen's Kappa = 0.9100
- ROC-AUC = 0.9956

The classification performance is highly effective across all three classes.

- Bipolar II: Precision = 0.91, Recall = 0.91, F1score = 0.91
- Unipolar: Precision = 0.91, Recall = 0.91, F1score = 0.91
- Bipolar I: Precision = 1.00, Recall = 1.00, F1score = 1.00

Although the performance is strong, it is slightly lower than the CatBoost-based hybrid model.

Thus, the CatBoost + Shared NN model outperforms the LightGBM + Shared NN model in overall classification performance. The performance of both

hybrid models was compared using graphical visualization across multiple evaluation metrics.

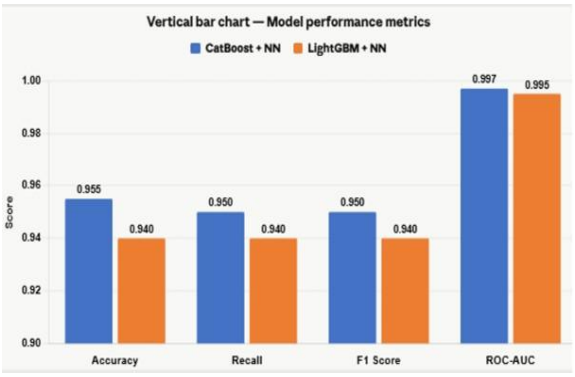

Fig. 4 Overall model performance comparison: accuracy, recall, F1-score, and ROC-AUC

Fig. 4 compares the performance metrics like accuracy, recall, F1-score and ROC-AUC. The CatBoost + Shared NN model shows consistently better performance. From the graph, it is clearly observed that the CatBoost-based hybrid model provides slightly superior performance across all metrics.

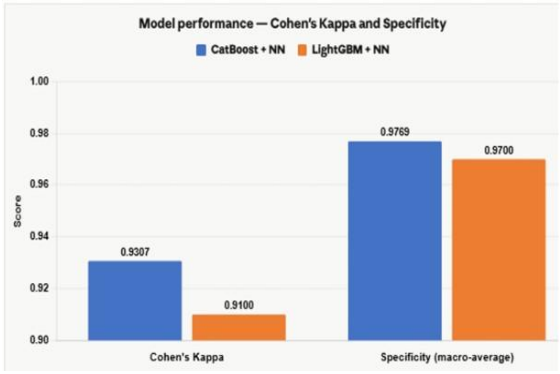

Fig. 5 Overall model performance comparison: Cohen's kappa and specificity

Fig. 5 compares the models using Cohen's Kappa and specificity, which are important in clinical classification systems. The CatBoost model achieved:

- Cohen's Kappa = 0.9307
- Average Specificity = 97.69%

The LightGBM model achieved:

- Cohen's Kappa = 0.9100
- Average Specificity = 97.00%

This indicates that the CatBoost-based model provides better agreement with actual clinical labels and better negative case identification.

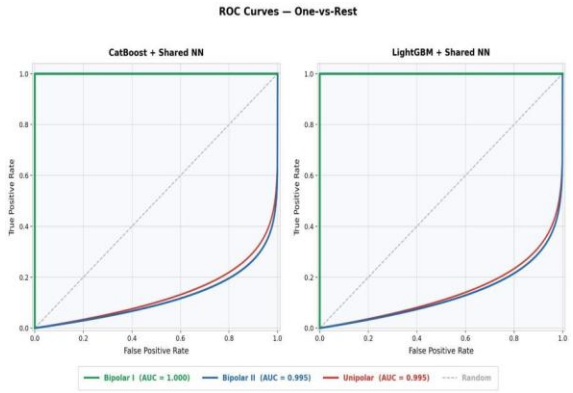

Fig. 6 ROC curve comparison of both hybrid models

Fig. 6 presents the one-vs-rest ROC curves for both hybrid models. The ROC curves for all three depression classes are positioned close to the top-left corner, indicating excellent class separability.

These results indicate that the proposed CatBoost-NN hybrid model provides superior classification performance for depression subtype prediction. The higher accuracy, Cohen's Kappa, and ROC-AUC values demonstrate its robustness in distinguishing clinically similar classes such as Bipolar II and Unipolar depression. The inclusion of SHAP-based explanations further improves the clinical reliability of the proposed system.

For both models:

- Bipolar I AUC = 1.000
- Bipolar II AUC ≈ 0.995
- Unipolar AUC ≈ 0.995

The CatBoost + Shared NN model shows a slightly superior curve shape and ROC-AUC value, confirming better classification capability.

The near-perfect ROC curves validate that both models are highly effective in distinguishing:

- Bipolar I
- Bipolar II
- Unipolar depression with CatBoost showing marginally better performance.

5.2 Streamlit-Based Clinical Prediction and Explainability Interface

To validate the practical applicability of the proposed system, A Streamlit-based clinical prediction and explainability interface was developed.

This interface allows clinicians or users to:

- select patient test records
- compare predictions from both hybrid models
- visualize important features
- understand model decisions using SHAP explanations

This improves model interpretability and enhances real-world clinical usability.

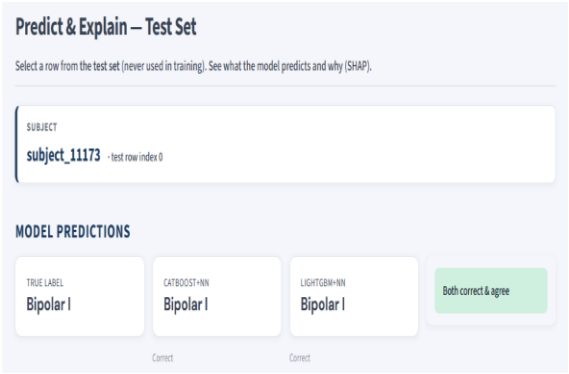

Fig. 7 Test set prediction comparison using hybrid models

Fig. 7 shows the prediction output for a selected test record. Both models:

- CatBoost + Shared NN
- LightGBM + Shared NN

correctly predicted the depression type as Bipolar I.

This prediction matches the actual ground-truth label in the test dataset, confirming the robustness of the hybrid models on unseen patient records.

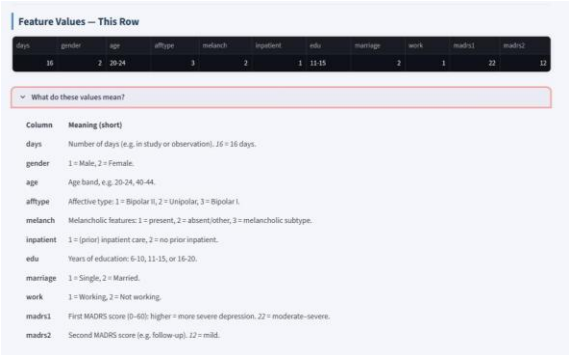

Fig. 8 Clinical feature values and patient details

Fig. 8 shows the clinical feature values of the patient, including:

- age group
- gender
- education level
- employment status
- inpatient status
- melancholic features
- episode duration
- MADRS severity scores

This panel provides a clear clinical interpretation of the input values used for prediction.

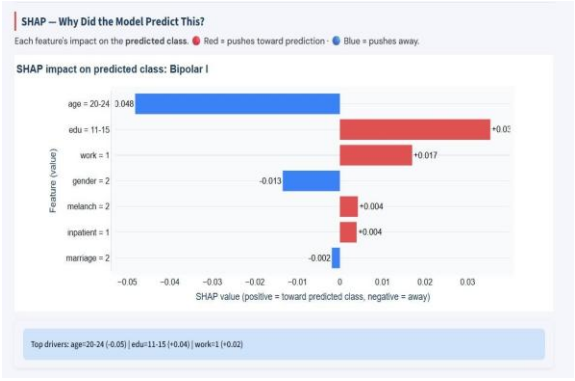

Fig. 9 SHAP-based feature contribution visualization

Fig. 9 presents the SHAP feature contribution graph. This visualization explains the reason behind the predicted class.

- Red bars - features pushing the prediction toward Bipolar I • Blue bars - features pushing away from Bipolar I

Important features such as:

- education level
- work status
- MADRS scores

show stronger contribution toward the final prediction. This significantly improves explainability and clinician trust.

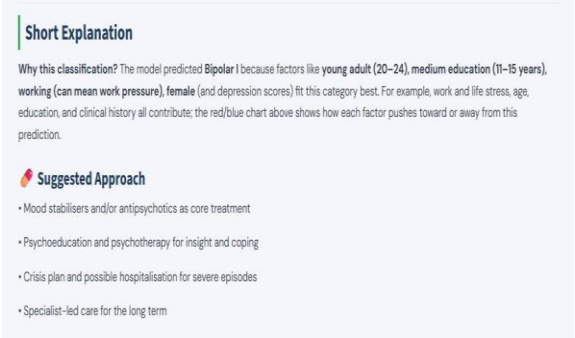

Fig. 10 Text explanation and clinical interpretation output

Fig. 10 shows the automatically generated textual explanation for the predicted depression class. The explanation summarizes:

- important clinical factors
- severity indicators
- demographic influence
- model reasoning in simple language.

This makes the prediction understandable even for nontechnical medical professionals and supports clinical decision-making.

5.3 Comparison with Other Related Works

Unlike previous studies that mainly focus on binary depression detection, the proposed framework performs multi-class subtype classification while maintaining high accuracy and explainability, making it more suitable for clinical decision support systems.

Although the proposed framework achieved high classification performance, the study is limited by the use of a relatively small publicly available dataset consisting of 55 participants. The use of window-based feature extraction increased the number of samples; however, these samples are derived from the same individuals and may not fully reflect population-level variability. Future work should validate the proposed model on larger multi-center datasets to improve generalizability and clinical reliability.

Table 2 represents the comparison of methodologies used, dataset, type of classification and accuracy for various papers with respect to the proposed implementation.

Table 2. Comparison with other related papers

Study	Method	Dataset	Task	Accuracy
Patil et al. [1]	RF + ANN	Motor activity	Binary	80.0%
Choi et al. [3]	XGBoost	Actigraphy	Binary	~85%
Huang et al. [4]	XGBoost + SHAP	Postpartum depression dataset (1065 women)	Binary	95.0%
Zhang et al. [13]	SVM/RF/KNN	Motor signals	Binary	~82%
Proposed	CatBoost-NN	Depresjon	Multiclass	95.38%
Proposed	LightGBMNN	Depresjon	Multiclass	94.00%

VI. CONCLUSION

The work “Intelligent Depression Classification System based on hybrid models using Actigraphy” aims to assess mental health related disorders automatically. This system uses wearable actigraphy data, hybrid machine learning models and explainable artificial intelligence to provide accurate depression type. By using minute-level data collected continuously from wearable devices, the system ensures continuous monitoring of patient’s daily activities, sleep patterns, and circadian rhythm, which are related to depression types.

This system uses tree-based models such as CatBoost and LightGBM with Artificial Neural Networks, which clearly classifies the depression type. The tree-based models learn from clinical data, while the neural network learns from non-linear relationships in the actigraphy data. This hybrid learning approach improves the overall prediction compared to traditional single-model methods.

These models can classify people into Bipolar I, Bipolar II, and Unipolar depression categories with high performance. From the results, the CatBoost + Neural Network model has an accuracy of 95.38% than the LightGBM + Neural Network model with 94.00% accuracy. The classification reports, confusion matrices and ROC-AUC scores confirm the effectiveness of the proposed hybrid models in distinguishing different depression classes with high precision, recall, and F1-score values. In addition to performance, the use of Explainable AI techniques such as SHAP helps explain the model predictions by identifying the most important features that result in depression detection.

Overall, this intelligent depression classification system indicates the use of wearable data, hybrid machine learning models, and explainability

techniques which lead to more accurate and clinically meaningful solutions for automated depression type classification. This approach reduces dependence on traditional methods and helps in early detection and better mental health monitoring.

VII. DECLARATION

Funding The authors did not receive support from any organization for the submitted work. No funding was received for conducting this study.

Competing interests The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval This study used a publicly available, fully anonymized dataset (the Depresjon dataset, originally published by Simula Research Laboratory and available on Kaggle). No human participants were directly recruited, and no personally identifiable information was used. The use of this publicly available dataset does not require ethics committee approval under applicable institutional and national guidelines.

Consent to participate Not applicable. This study used a pre-existing, anonymized, publicly available dataset. No direct participant recruitment was carried out.

Consent to publish Not applicable.

Data availability The Depresjon dataset used in this study is publicly available on Kaggle, originally published by the Simula Research Laboratory. It can be accessed at: https://www.kaggle.com/datasets/arashnic/the-depression-dataset. No additional data were generated or collected by the authors.

Author contributions Conceptualization: Thamina Anzum A, Dr. Thirumahal R, Vinithaa P, Gobika R, Naveen Ragav K, Naveen P;

Methodology: Thamina Anzum A, Vinithaa P, Gobika R, Naveen Ragav K, Naveen P;

Formal analysis and investigation: Dr. Thirumahal R

Writing – original draft preparation: Thamina Anzum A;

Writing – review and editing: Thamina Anzum A, Vinithaa P, Gobika R, Naveen Ragav K, Naveen P;

Supervision: Dr. Thirumahal R.

All authors read and approved the final manuscript.

REFERENCES

[1] Patil A, Shah D, Shah A, Gala M (2023) A hybrid approach for depression classification: random forest-ANN ensemble on motor activity signals. arXiv:2310.09277

[2] Dong Y, Wen H, Lu C et al (2025) Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants. BMC Psychiatry 25:883

[3] Choi JG, Ko I, Han S (2021) Depression level classification using machine learning classifiers based on actigraphy data. IEEE Access 9:116622–116646

[4] Abir RA, Chowdhury M, Zeba LT, Karib N, Arafat MY, Akter A (2025) Postpartum psychiatric disorder prediction using machine learning and explainable AI. In: Proceedings of the 9th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, pp 1–6

[5] Aleem S, ul Huda N, Amin R, Khalid S, Alshamrani SS, Alshehri A (2022) Machine learning algorithms for depression: diagnosis, insights, and research directions. Electronics 11(7):1111

[6] Shah SM, Gillani SA, Baig MSA, Saleem MA, Siddiqui MH (2025) Advancing depression detection on social media platforms through fine-tuned large language models. Online Social Networks and Media 46:100311

[7] Kerasiotis M, Ilias L, Askounis D (2024) Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining 14:196

[8] Khan S, Saeed SMU, Frnda J, Arsalan A, Amin R, Gantassi R et al (2024) A machine learning based depression screening framework using temporal domain features of electroencephalography signals. PLoS ONE 19(3):e0299127

[9] Jamali AA, Berger C, Spiteri RJ (2024) Identification of depression predictors from standard health surveys using machine learning. Current Research in

Behavioral Sciences 7:100157

[10] Vysotska V, Bychkov I, Lynnyk R, Fedchuk R, Lozynska O, Markiv O (2025) MindGuard: intelligent system for depressive and suicidal intentions detection in social media texts based on CNN, BERT and RoBERTa. In: Proceedings of the IEEE KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, pp 1–6

[11] Kim JW, Lee T, Lim B, Park SH, Park JH, Jeong I, Park K, Kang HJ, Jeon E, Kim SW, Jhon M, Lee H, Kim JM (2026) Actigraphy-based step analysis for the detection of depressed mood: an explainable machine learning approach. Journal of Affective Disorders 392:120104

[12] Price GD, Heinz MV, Collins AC, Jacobson NC (2024) Detecting major depressive disorder presence using passively-collected wearable movement data in a nationally-representative sample. Psychiatry Research 332:115693

[13] Schulte A, Breiksch T, Brockmann J, Bauer N (2022) Machine learning based classification of depression using motor activity data and autoregressive model. In: Proceedings of GMDS, pp 25–32

[14] Ghate R, Kalnad N, Walambe R, Kotecha K (2023) Transfer learning for real-time deployment of a screening tool for depression detection using actigraphy. arXiv:2303.07847

[15] Vishnawa V, Verma NK (2025) A systematic review of AI and multimodal techniques for depression detection and AR/VR-based therapy. In: Proceedings of the 8th International Conference on Emerging Technologies in Computer Engineering: Advances in Computing, Healthcare and Smart Systems (ICETCE), Jaipur, India, pp 190–197

[16] Kim S, Imieye O, Yin Y (2025) Interpretable depression detection from social media text using LLM-derived embeddings. arXiv:2506.06616

[17] Kuzmin G, Strepetov P, Stankevich M, Shelmanov A, Smirnov I (2024) Mental disorders detection in the era of large language models. arXiv preprint

[18] Chen Z, Yang R, Fu S, Zong N, Liu H, Huang M (2023) Detecting Reddit users with depression using a hybrid neural network SBERT-CNN. In: Proceedings of the IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, USA, pp 193–199

[19] Farruque N, Goebel R, Sivapalan S, Zaïane OR (2024) Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. Language Resources and Evaluation 58(3):1013–1041

[20] Baydili İ, Tasci B, Tasci G (2025) Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. Behavioral Sciences 15(3):352