JOURNAL OF DATA SCIENCE | Vol.2025:01
eISSN:2805-5160

Enhancing Bipolar Disorder Detection Using Heterogeneous Ensemble Machine Learning Techniques

Lingeswari Sivagnanam[1], N. Karthikeyani Visalakshi[2]

[1,2]Department of Computer Science, Government Arts and Science College, Kangeyam, India

Email: tejumitu1279@gmail.com[1], karthichitru@yahoo.co.in[2]

Abstract

This paper introduces a novel Heterogeneous Ensemble Machine Learning (HEML) approach designed to detect bipolar disorder, a significant healthcare challenge that demands precise and prompt diagnosis for effective treatment. The HEML method integrates multiple machines learning models, incorporating various physiological, behavioral, and contextual data from patients. By using a comprehensive feature selection technique, relevant features are extracted from each data source and utilized to train individual classifiers for detecting mental disorders. The classifiers include Adaboost, Decision Tree, K-nearest neighbors, Multilayer Perceptron, Random Forest, Relevance Vector Machine, and XGB, with Logistic Regression serving as the meta-model. This ensemble of classifiers enhances overall performance by capturing a wider range of characteristics related to mental disorders. The research evaluates the HEML method across three bipolar disorder datasets: Dataset1 (a multimodal dataset), Dataset2 (a sensor-based dataset), and Dataset3 (a real-time dataset). The HEML approach surpasses traditional methods, achieving superior accuracy rates of 95.21% with Dataset 1, 99.28% with Dataset 2, and 99% with Dataset 3. It outperforms individual models in detecting bipolar disorder, delivering the best Precision, Recall, F1 score, and Kappa Score. This comparative analysis advances the field of mental health diagnosis by leveraging the strengths of ensemble machine learning to improve accuracy and reliability in detection methods.

Keywords

Bipolar disorder, HEML, RVM, Random forest, XGB boost

Introduction

A severe and perpetual chronic mental illness, Bipolar Disorder (BD), often emerges in early adolescence (18.4–20 years) and contributes to 6.8% (4.9–9.1) of years of life modified for impairment due to psychological illness. Untreated BD can lead to heightened depression, hypermania, and suicidal ideation, as 10-20% of BD patients attempt suicide (Müller-Oerlinghausen et al., 2002). Detecting BD at an early stage and providing timely therapy is crucial to prevent such severe outcomes.

Submission: 27 July 2024; **Acceptance:** 24 April 2025; **Available Online:** April 2025

Scientific research has increasingly focused on BD detection over therapeutic techniques. One approach involves examining genetic predisposition by identifying BD risk among first-degree relatives, with an estimated 4.2–22.4% transition rate (Hajek et al., 2013). Risk analysis tools also assess clinical factors to aid diagnosis, including structured interviews like the Bipolar At-Risk States Revised and the Semistructured Interview of Bipolar At-Risk States. Additional BD risk assessment tools, such as the Early Phase Inventory for Bipolar Disorders and the Bipolar Prodrome Symptom Scale-Prospective, utilize DSM-IV criteria to evaluate diverse risk factors. These interviews, combined with biomarkers like structural neuroimaging, facilitate early detection and effective healthcare delivery for BD (Arnone et al., 2009).

Various physiological anomalies are identifiable in BD through biomarker research, such as reduced cortical thickness in frontal, parietal, and temporal brain regions and smaller subcortical structures like the hippocampus, thalamus, and amygdala (Mathew et al., 2014). Fine-grained parcellation methods have enabled a more detailed examination of these subregions, with studies showing hippocampal volume reduction in BD. For instance, a large-scale analysis of 4,698 patients identified lower levels in specific hippocampal subregions, with some variations observed in BD and schizophrenia (SZ) patients (Brown, 2010; Maity et al., 2022).

Machine learning (ML) classifiers have shown promise in assessing BD, enabling early detection, diagnosis, and prediction of mental and physical illnesses (Ganasigamony & Selvaraj, 2022). Ensemble learning models, known for their high accuracy, outperform individual ML models by combining multiple predictions, which reduces error and increases robustness (Dwyer et al., 2018; Ali et al., 2019). Ensemble classifiers address the limitations of single models by reducing variance (through bagging) and bias (through boosting), resulting in improved performance (Fitriyani et al., 2019).

The stacking ensemble method, introduced by Wolpert (1992), further enhances model accuracy by combining predictions from different classifiers through a meta-model. In this approach, a separate model generates predictions based on outputs from other models, leveraging each classifier's strengths for better outcomes. Research indicates that stacking often outperforms single classifiers across diverse datasets, making it an effective technique for BD detection (Wan et al., 2017; Rotenberg et al., 2021).

BD's extensive emotional fluctuations—ranging from depressive to manic states—affect millions globally, with significant social, personal, and economic consequences. Early detection remains challenging due to BD’s complex and heterogeneous nature, varied symptoms, and comorbidities (Sivagnanam & Visalakshi, 2023).

Conventional approaches relying on medical evaluations are time-consuming and prone to inaccuracies. ML systems, predominantly designed for homogeneous data, struggle to capture BD’s complexities, necessitating advanced methods to integrate diverse data types for precise diagnosis (Rao et al., 2020). To address these challenges, innovative technologies leveraging heterogeneous ensemble models can optimize BD diagnosis accuracy (Mateo-Sotos et al., 2022). This methodology will apply a Heterogeneous Ensemble approach, combining various classifiers through stacking, with logistic regression as the meta-model, to enhance BD detection accuracy (Peerbasha & Surputheen, 2021; Fonseca et al., 2018; Achalia et al., 2020).

Methodology

Bipolar disorder is an irrelevant mental health condition having up and down mood swings, and emotional stress every day. It is very crucial to detect at an early stage and provide efficient treatment. In this research, an ensemble heterogeneous classifier is introduced for detecting bipolar disorder precisely. This system incorporates multiple machine-learning techniques to achieve optimal prediction accuracy and resilience.

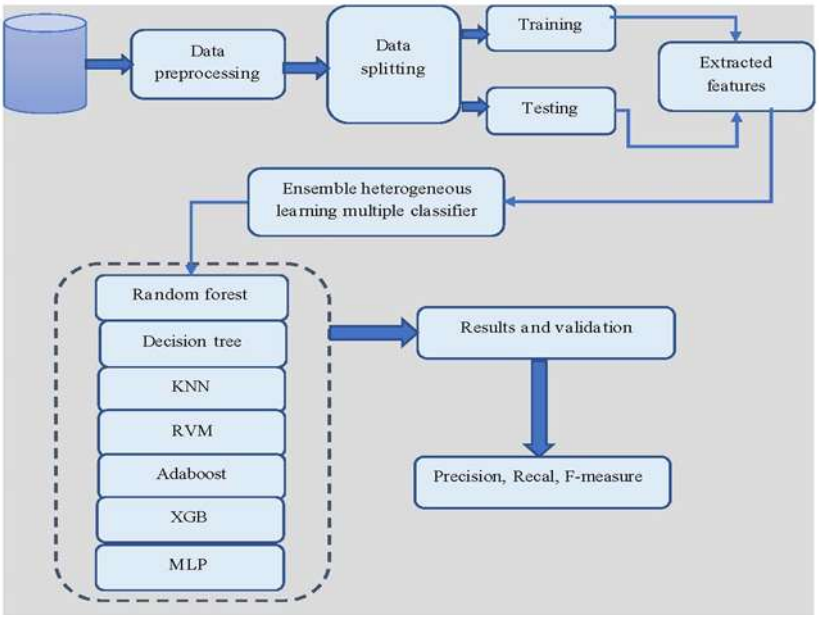

Figure 1. Proposed HEML architecture

Figure 1 shows the proposed HEML framework as an architecture representation of the contribution applied in this work.

Data Collection

In this model, data is collected from various means such as patient records, Electronic Health Records (EHRs), smart devices, and medical consultations. The crucial characteristics required for this system to evaluate the comprehensive perspective of the patient are the patient's medical history, mood swings rate, activity range, demographic data, sleep routine, medication compliance and speech mode.

Dataset-1

Dataset1 consists of transcriptions from psychotherapy interviews conducted on the Google Cloud platform. The Audio/Visual Emotion Challenge (AVEC) [12] utilizes audio and visual data to analyze human emotional states and behaviors. AVEC provides a platform for researchers to test their systems on standardized datasets, advancing emotion recognition and related fields. The dataset includes recorded sessions with various emotional states, labeled to indicate emotional conditions or behaviors, along with synchronized audio and video recordings.

Dataset-2

Dataset2, known as the Wearable Stress and Affect Detection (WESAD) dataset, is publicly available for stress detection using data from wrist-worn devices. This dataset includes physiological signals recorded during various emotional states (stress, neutral, amusement). It offers preprocessed data for detecting patients' normal and abnormal conditions based on device-monitored data, making it useful for mental health research [13].

Dataset-3

Dataset3 comprises records collected from a department with a high number of recent patient samples. It includes patient histories and clinical details stored in a .CSV file. Features collected include age, gender, sleepiness, aggression, and depression. Data reduction techniques were applied to compress the dataset while maintaining analytical accuracy. Dimensionality reduction methods, such as label encoding and one-hot encoding, were used to simplify the data. Statistical significance was used to determine the best and worst attributes. Preprocessing addressed class imbalance and outlier removal using ensemble learning methods. The Variational Autoencoder Synthetic Minority Over-sampling Technique (VAE SMOTE) was employed to balance the dataset by generating synthetic samples, addressing imbalanced data common in fraud detection, medical diagnosis, and anomaly detection [14][15]. Performance evaluation of the model ensured the effectiveness of the SMOTE technique.

Data Preprocessing

The procedures involved in data preprocessing are cleaning, assembling and encoding the data. An imputation method is utilized to operate the lost data. The consistent nature is maintained among the attributes by setting a certain limit for the arithmetic values. This system employs one-hot encoding or label encoding to convert the categorized data into arithmetic values. Moreover, noise deduction method like Principal Component Analysis (PCA) or Independent Component Analysis (ICA) is utilized to optimize the quality of data.

Data Splitting

To ensure that the model can be examined with unfamiliar data, the preprocessed data is separated into training and testing datasets in the ratio of 80-20 or 70-30. One type of cross validation technique is k-fold cross-validation which is applied to optimize the model’s robustness as well as prevent overfitting. In this module, split the training data into k subsets and training the system in k times. The validation set is chosen various subsets periodically.

Feature Extraction

The optimal model performance and reduced complexity is achieved through feature extraction. Time-series analysis is employed to extract attributes from sleep routine, heart rate deviation and activity range. Through the mood and activity levels, statistical features such as mean, standard deviation, skewness and kurtosis can be extracted. Regarding behavioral feature extraction, the social behavior, daily routine and mood rating patterns is analyzed. Natural language processing (NLP) method extract feature from the medical history and consulting notes along with subject simulation and emotional analysis.

Training

The training module begin with choosing an ensemble of multiple machine learning methods and training them. Random forest, Decision Tree, K-Nearest Neighbours (KNN), Adaboost, Relevant Vector Machine (RVM), Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP) models involved in this system. The various data pattern is captured by every model to optimize the ensemble's overall stability. The optimal model performance is achieved through hyper parameter tuning by using models like grid search or random search.

Testing

To evaluate the accuracy, F-measure, recall and precision, the trained models are analyzed in the testing sets. This procedure assure that make the model handle fresh and unknown data effectively. Emphasize the section which require optimization and detailed information of model's efficacy is analyzed from the performance measures.

Ensemble Heterogeneous Learning

By using methods like majority voting, weighted averaging, or stacking in ensemble heterogeneous classifier, compiles the predictions from the all models. This model utilize each classifier's potential to improve the system's entire prediction accuracy. A meta-learner such as a logistic regression model is employed in stacking to compile the base models prediction and optimize the ensemble's performance.

Validation

By comparing the predicted value with real values, an ensemble model's compiled findings are examined with real facts which assures the stability of model's prediction. The model's overall accuracy and performance is demonstrated through performance measures like precision, recall, and F-measure. The proper diagnosis and management is eventually achieved by incorporating multiple classifier in this approach which predicts the bipolar disorder precisely.

Algorithm for HEML

1. Initialize a list of diverse classifiers (e.g., Decision Tree, RVM, Random Forest, XGBoost)
2. Train the classifier on the training dataset
3. Store the trained classifier in a model list
4. Use the trained classifier to make predictions on the test dataset
5. Collect the predictions from each classifier
6. Combine Prediction
7. Aggregate the predictions from all classifiers (e.g., majority voting, weighted voting, or averaging)
8. Evaluate Performance
9. Compare the combined predictions against the ground truth labels
10. Calculate performance metrics (e.g., accuracy, precision, recall, F1 score)
11. Output the combined predictions and performance metrics

Results and Discussion

Confusion matrix

In the confusion matrix, the prediction output of classification error is represented specifically. It is of two prediction classes: correct or incorrect prediction. The broken down and counted data determines the correct or incorrect prediction outcome. The best part of it is that it evaluates the type of error made by that particular classifier as well as the way of making mistakes.

Description of the Terms:

- Positive (P): Observation is positive
- Negative (N): Observation is not positive
- True Positive (TP): Observation is positive and is predicted to be positive.
- False Negative (FN): Observation is positive but is predicted negative.
- True Negative (TN): Observation is negative and is predicted to be negative.
- False Positive (FP): Observation is negative but is predicted positive.

Recall

The Recall is evaluated by the division of properly categorized samples by the overall positive examples. The lower FN value and high recall values shows that samples are properly identified.

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

Precision

Precision is calculated by dividing the overall properly classified positive samples by the total predicted positive examples. The positive output is denoted by the high precision result. It is shown below

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

F-measure

F-measure is calculated by Recall and precision result. Rather than arithmetic mean-measure is evaluated by harmonic mean which employs more efficient with higher values. Consider these, F-measure is usually less than the Recall or precision. It is evaluated as follows.

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

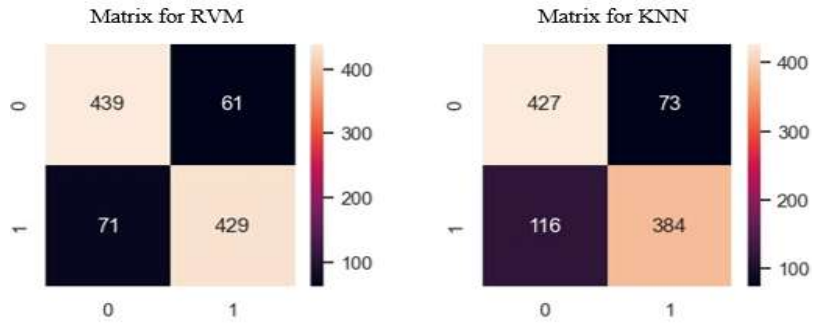

Figure 2. Matrix RVM and KNN

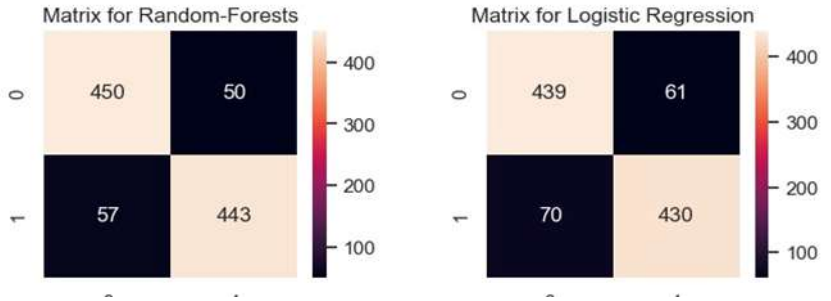

Figure 3. Matrix Random-Forests and Logistic Regression

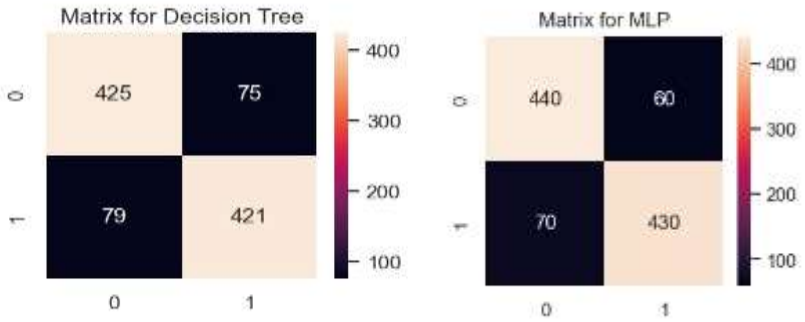

Figure 4. Matrix Decision tree and MLP

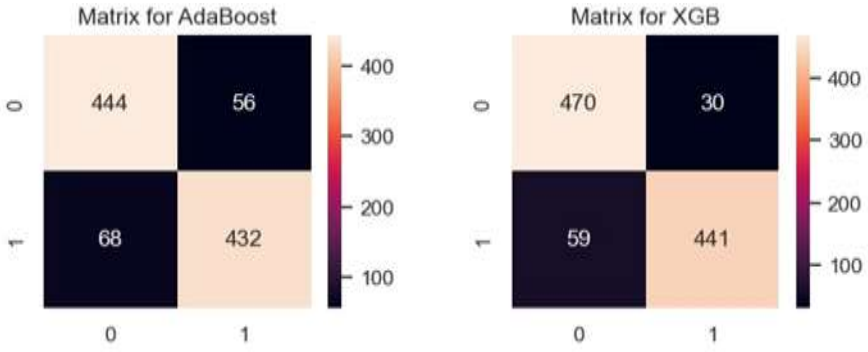

Figure 5. Matrix AdaBoost and XGB

The above figures depict the confusion matrix of the classifiers applied in this paper, thus showing a clear view of the work.

Table 1. Performance metrics of individual classifiers on Dataset1

Classifier	Accuracy	Precision	Recall	F1-score	Cohen Kappa Score
Ada boost	91.1	91.09	91.47	90.86	9.36
Decision Tree	80.71	82.02	20.91	81.15	79.16
KNN	90.71	90.55	91.83	90.67	89.90
MLP	95	95.26	95.14	95.02	94.60
Random Forest	90.71	90.63	91.32	90.53	89.98
RVM	88.92	90.62	90.11	89.56	88.06
XGB	92.14	92.01	92.52	92.09	91.52

HEML (proposed)	95.71	96.03	95.81	95.85	95.37

Table-1 shows the results of the individual classifiers with their performance metrics using dataset1 (Multimodal data) with its Accuracy, Precision, Recall, F1-score, Cohen kappa score.

Table 2. Performance metrics of individual classifiers on Dataset2

Classifier	Accuracy	Precision	Recall	F1-score	Cohen Kappa Score
Ada boost	96.84	96.96	96.84	96.78	95.58
Decision Tree	92.53	92.64	92.53	92.49	89.54
KNN	93.25	93.24	93.25	93.21	90.50
MLP	80.02	78.88	80.02	79.51	71.34
Random Forest	95.44	95.56	95.44	95.10	93.57
RVM	73.04	74.99	73.00	72.89	60.95
XGB	95.02	95.28	95.17	94.98	93.99
HEML (proposed)	96.84	96.96	96.84	96.78	95.58

Table-2 shows the results of the individual classifiers with their performance metrics using dataset2 (Sensor data) with its Accuracy, Precision, Recall, F1-score, Cohen kappa score.

Table 3. Performance metrics of individual classifiers on Dataset3

Classifier	Accuracy	Precision	Recall	F1-score	Cohen Kappa Score
Ada boost	95	95.14	95	95.01	89.90
Decision Tree	96.25	96.54	96.25	96.26	92.45
KNN	92.5	93.59	92.5	92.58	85.04
MLP	95	95.51	95	95.01	89.96
Random Forest	92.5	93.59	92.5	92.52	85.04
RVM	96.5	96.36	96.45	96.49	93.85
XGB	97.5	97.63	97.50	97.50	94.95
HEML (proposed)	99	99.2	97.77	98.87	97.46

Table-3 shows the results of the individual classifiers with their performance metrics using dataset3 (Real data) with its Accuracy, Precision, Recall, F1-score, Cohen kappa score.

Table 4. Comparison between the existing classifier

Classifier	Accuracy
XGB [13]	94%
SVM [16]	87%
Random forest [17]	74%
HEML [Proposed work]	99%

In table 4, the different classifier's prediction result is tabulated to identify the bipolar disorder as well as illustrates the performance of proposed ensemble heterogeneous classifier (HEML) with the existing algorithms. In data science technology, Extreme Gradient Boosting (XGB) classifier is the most prevalence method since it has superior efficiency and performance

especially in managing large data sets and intricated patterns. XGB achieved accuracy of about 94%. Support Vector Machine (SVM) performs better in multidimensional spaces and handling classification challenges but it reaches an accuracy of 87% which is less than the existing XGB and proposed HEML model. The Random Forest classifier utilizes various decision tree for optimizing prediction accuracy but acquires an accuracy of 74%. Even though it poses more durability and interpretability, its performance in this case is extremely lower than XGB and proposed HEML.

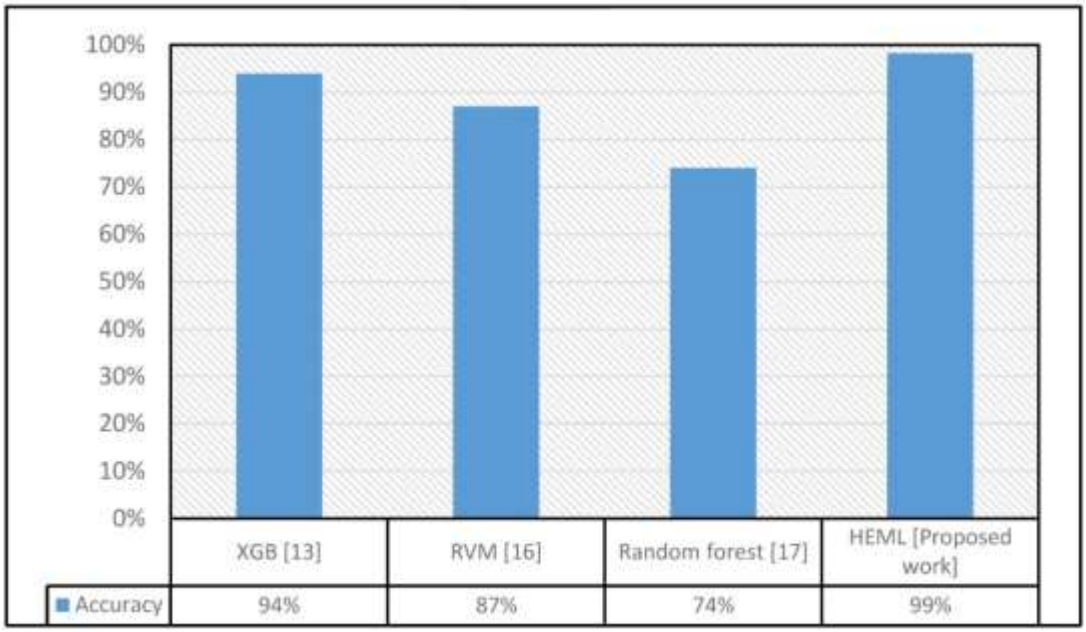

Figure 6. Comparison between the existing algorithms

In figure 6, a comparative analysis done between the prediction accuracy of multiple classifiers to predict the bipolar disorder. Random forest classifier achieved a lowest accuracy of all other models of about 74%. Support Vector Machine (RVM) performs better than random forest which has an accuracy of about 87%. High performance is achieved by Extreme Gradient Boosting (XGB) with an accuracy of 94%. Our proposed ensemble heterogeneous model (HEML) performed better than all other models with an accuracy of 99% which is having great potential for bipolar disorder's prediction and diagnosis.

Table-5 different data sets compared with the proposed classifier

Datasets	Proposed Accuracy	Classifier
Dataset 1	HEML	98.5%
Dataset 2	HEML	99%
Dataset 3	HEML	99%

Table-5 describes about an accuracy of proposed HEML classifier is obtained for three various data sets is represented in Table 2.A HEML classifier reached an accuracy of about 95% for first data set. The accuracy has been improved more for both dataset 2 & dataset 3 is about

99%.Hence, the outstanding performance of classifier is demonstrated from this outcome. Also it can handle the multiple datasets effectively and offer persistent accuracy in all circumstances.

Acknowledgments

I thank Dr. Karthikeyani Visalakshi for her contributions to this work.

References

Achalia, R., Sinha, A., Jacob, A., Achalia, G., Kaginalkar, V., Venkatasubramanian, G., & Rao, N. P. (2020). A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. *Asian Journal of Psychiatry*, *50*, 101984. https://doi.org/10.1016/j.ajp.2020.101984

Ali, R., Hardie, R. C., Narayanan, B. N., & De Silva, S. (2019, July 15–19). Deep learning ensemble methods for skin lesion analysis towards melanoma detection. In *Proceedings of the 2019 IEEE National Aerospace and Electronics Conference (NAECON)* (pp. 311–316). Dayton, OH, USA. http://dx.doi.org/10.1109/NAECON46414.2019.9058245

Arnone, D., Cavanagh, J., Gerber, D., Lawrie, S. M., Ebmeier, K. P., & McIntosh, A. M. (2009). Magnetic resonance imaging studies in bipolar disorder and schizophrenia: Meta-analysis. *British Journal of Psychiatry*, *195*(3), 194–201. https://doi.org/10.1192/bjp.bp.108.059717

Brown, G. (2010). Ensemble learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 312–320). Springer. https://doi.org/10.1007/978-0-387-30164-8_252

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Fitriyani, N. L., Syafrudin, M., Al_an, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access*, *7*, 144777–144789. https://doi.org/10.1109/ACCESS.2019.2945129

Fonseca, M., Andrades, R., Bach, S., Wiener, C., & Oses, J. (2018). Bipolar and schizophrenia disorders diagnosis using artificial neural network. *Neuroscience and Medicine, 9*(4), 209–220. https://doi.org/10.4236/nm.2018.94021

Ganasigamony, W. J., & Selvaraj, M. A. A. (2022). Computer assisted diagnosis of bipolar disorder using invariant features. *Concurrency and Computation: Practice and Experience*, e6984. https://doi.org/10.1002/cpe.6984

Hajek, T., Cullis, J., Novak, T., Kopecek, M., Blagdon, R., Propper, L., Stopkova, P., Duffy, A., Hoschl, C., Uher, R., et al. (2013). Brain structural signature of familial predisposition for

bipolar disorder: Replicable evidence for involvement of the right inferior frontal gyrus. *Biological Psychiatry, 73*(2), 144–152. https://doi.org/10.1016/j.biopsych.2012.06.015

Mateo-Sotos, J., Torres, A. M., & Santos, J. L. (2022). A machine learning-based method to identify bipolar disorder patients. *Circuits, Systems, and Signal Processing, 41*, 2244–2265. https://doi.org/10.1007/s00034-021-01889-1

Luján, M. Á., Torres, A. M., Borja, A. L., Santos, J. L., & Mateo Sotos, J. (2022). High-precise bipolar disorder detection by using radial basis functions-based neural network. *Electronics, 11*(3), 343. https://doi.org/10.3390/electronics11030343

Maity, S., Mandal, R. P., Bhattacharjee, S., & Chatterjee, S. (2022). Variational autoencoder-based imbalanced Alzheimer detection using brain MRI images. In L. Mandal, J. M. R. S. Tavares, & V. E. Balas (Eds.), *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing*. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-19-1657-1_14

Mathew, I., Gardin, T. M., Tandon, N., Eack, S., Francis, A. N., Seidman, L. J., Clementz, B., Pearlson, G. D., Sweeney, J. A., & Tamminga, C. A. (2014). Medial temporal lobe structures and hippocampal subfields in psychotic disorders: Findings from the bipolar-schizophrenia network on intermediate phenotypes (B-SNIP) study. *JAMA Psychiatry, 71*(7), 769–777. https://doi.org/10.1001/jamapsychiatry.2014.453

Müller-Oerlinghausen, B., Berghöfer, A., & Bauer, M. (2002). Bipolar disorder. *Lancet, 359*(9302), 241–247. https://doi.org/10.1016/S0140-6736(02)07450-0

Peerbasha, S., & Surputheen, M. M. (2021). A predictive model to identify possible affected bipolar disorder students using Naive Bayes, Random Forest, and RVM machine learning techniques of data mining and building a sequential deep learning model using Keras. *International Journal of Scientific and Technology Research, 21*(5), 267–274. https://doi.org/10.22937/IJCSNS.2021.21.5.36

Rao, G., Peng, C., Zhang, L., Wang, X., & Feng, Z. (2020). A knowledge-enhanced ensemble learning model for mental disorder detection on social media. In G. Li, H. Shen, Y. Yuan, X. Wang, H. Liu, & X. Zhao (Eds.), *Knowledge Science, Engineering and Management. KSEM 2020. Lecture Notes in Computer Science* (Vol. 12275). Springer, Cham. https://doi.org/10.1007/978-3-030-55393-7_17

Rotenberg, L. S., Borges-Júnior, R. G., Lafer, B., Salvini, R., & Dias, R. D. S. (2021). Exploring machine learning to predict depressive relapses of bipolar disorder patients. *Journal of Affective Disorders, 295*, 681–687. https://doi.org/10.1016/j.jad.2021.08.127

Sivagnanam, L., & Visalakshi, N. K. (2023). Detection of bipolar disorder by means of ensemble machine learning classifier. *Data and Metadata, 2*, 134–134. https://doi.org/10.56294/dm2023134

Wan, Z., Zhang, Y., & He, H. (2017, November). Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE Symposium Series on Computational*

Intelligence (SSCI) (pp. 1–7). Honolulu, HI, USA. https://doi.org/10.1109/SSCI.2017.8285168

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1