Optimized Machine Learning Models For Early Detection Of Alcohol Use Disorder: A Hybrid Approach

Dr. B. Santhosh Kumar[1], Dr. D. Kumaresan[2*]

[1]Assistant Professor, Department of Computer Applications, Periyar Arts College, Cuddalore, Tamil Nadu, India.
[2]Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India.*

Email: [1]santhoshcdm@gmail.com
Corresponding Email: [2]aucsedks@yahoo.co.in*

KEYWORDS

Alcohol Use Disorder (AUD), Machine Learning Optimization, Metaheuristic Algorithms, Deep Learning Models, Explainable AI (XAI), And Clinical Decision Support Systems.

ABSTRACT:

Alcohol Use Disorder (AUD) remains a critical global health issue, necessitating the development of advanced diagnostic systems for early and accurate detection. Conventional diagnostic methods often exhibit subjectivity and limited predictive capability, emphasizing the need for intelligent computational techniques. This research introduces a hybrid optimization framework that integrates machine learning models with metaheuristic optimization strategies to enhance AUD detection. By incorporating evolutionary algorithms, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), alongside deep learning techniques, the proposed approach optimally selects features, fine-tunes hyperparameters, and reduces overfitting. A diverse dataset combining clinical, behavioral, and neuroimaging data is used to train and validate the model, ensuring broad applicability across different populations. Comparative evaluations with traditional machine learning models indicate that the hybrid-optimized method substantially improves classification accuracy, sensitivity, and specificity in differentiating AUD from non-AUD cases. Additionally, explainable AI techniques are utilized to improve model interpretability, aiding healthcare professionals in understanding key predictive factors. The results highlight the potential of hybrid optimization in machine learning for AUD diagnosis, contributing to more reliable, data-driven clinical decision-making and early intervention strategies.

1. Introduction

Alcohol Use Disorder (AUD) remains a major global health issue, requiring the development of advanced predictive and diagnostic tools. Traditional assessment methods, such as self-reported questionnaires and clinical interviews, are often subjective and may lack accuracy. With the rise of artificial intelligence (AI) and machine learning (ML), researchers have explored data-driven approaches to enhance the identification and classification of AUD. Furthermore, the integration of metaheuristic optimization techniques within ML frameworks has been employed to improve model efficiency, optimize feature selection, and reduce overfitting. This section provides an overview of significant studies that have contributed to ML-based AUD detection, focusing on optimization methods, EEG-based classification, feature selection, and ensemble learning strategies.

Traditional diagnostic approaches for AUD are predominantly reliant on subjective clinical assessments and self-reported data, which are inherently vulnerable to inaccuracies and biases. This inherent subjectivity accentuates the urgent need for objective, data-driven methodologies that can facilitate earlier, more accurate detection and intervention. In recent years, Machine Learning (ML) and optimization techniques have gained prominence as powerful tools in this domain, offering the potential to analyze complex datasets and uncover patterns that might otherwise remain undetected through conventional diagnostic techniques.

This study introduces a hybrid optimization framework that synergistically integrates machine learning algorithms with metaheuristic optimization techniques to enhance the performance of AUD detection models. By utilizing evolutionary algorithms such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), in combination with advanced deep learning architectures, the proposed framework seeks to optimize feature selection, calibrate hyperparameters, and reduce the risks of overfitting. A diverse, multi-source dataset—comprising clinical, behavioral, and neuroimaging data—will be employed to train and validate the model, ensuring its robustness and generalizability across heterogeneous populations.

Several studies have highlighted the role of optimization techniques in enhancing ML-driven AUD detection. Zhang et al. (2018) introduced a classification model utilizing Cat Swarm Optimization (CSO) to improve accuracy, demonstrating that optimization algorithms enhance predictive performance by refining hyperparameters and feature selection. Similarly, Ebrahimi et al. (2023) developed a Decision Support System (DSS) incorporating ML models and clinical data, which improved early AUD detection and minimized subjective biases present in conventional diagnostic methods. Given its ability to capture neural activity patterns, EEG-based ML models have gained significant attention for AUD classification. Mumtaz et al. (2016) employed deep learning models to analyze EEG signals, achieving high accuracy in distinguishing individuals with AUD from non-affected subjects. Their findings align with those of Lopes et al. (2004), who applied wavelet transformation and artificial neural networks (ANNs) for event-related potential (ERP) classification, proving that wavelet-based feature extraction enhances diagnostic performance.

To further refine EEG-based classification, Ng et al. (2012) utilized Recurrence Quantification Analysis (RQA) to detect non-linear patterns in EEG signals, significantly improving classification stability compared to conventional time-domain techniques. Likewise, Ong et al. (2005) used Principal Component Analysis (PCA) to select relevant EEG channels, optimizing feature selection while minimizing computational complexity. Understanding the neurophysiological effects of AUD has been critical in designing accurate diagnostic frameworks. Michael et al. (1993) investigated EEG coherence patterns in individuals with AUD, revealing disrupted interhemispheric synchronization, which provided insights into alcohol-induced neurological impairments. From a behavioral perspective, Moss et al. (2007) conducted an extensive study to categorize subtypes of alcohol dependence, associating these subtypes with distinct neurophysiological and behavioral traits. Their research laid the groundwork for personalized treatment strategies tailored to specific AUD profiles. Weather dataset whose attributes represent the weather conditions and the class variable indicates whether the conditions are suitable for playing golf. Seven classification algorithms were used to measure accuracy, including J48, Random Tree, Decision Stump, Logistic Model Tree, Hoeffding Tree, Reduce Error Pruning, and Random Forest. Among these algorithms, Random Tree achieved the highest accuracy of 85.714% discussed by Rajesh and Karthikeyan (2017) Rajesh et al. (2019) explored Chronic disease data is analyzed with attributes representing topics, questions, data values, low and high confidence limits. Five classification algorithms are used to evaluate the data. The M5P decision tree approach is found to be the best algorithm for building a model compared to other decision tree approaches.

Feature selection is essential for optimizing ML models in AUD classification by identifying the most relevant variables while minimizing redundant data. Liao and Chin (2007) applied logistic regression models to disease classification using microarray data, emphasizing the importance of feature selection in handling high-dimensional datasets. Mamitsuka (2006) further refined feature selection techniques by implementing Receiver Operating Characteristic (ROC) curve-based methods, which significantly enhanced model performance by improving variable selection. Ensemble learning methods have been extensively explored to enhance predictive accuracy in AUD classification. Kuncheva and Rodríguez (2013) introduced a weighted voting ensemble framework, demonstrating superior performance compared to individual classifiers. Their study highlighted the advantages of integrating multiple ML models to increase classification robustness. The combination of multiple ML techniques has demonstrated promising results in improving AUD detection accuracy. Lopes et al. (2005) explored Learning Vector Quantization (LVQ) networks for ERP signal classification in at-risk individuals, validating LVQ's effectiveness in differentiating AUD-affected individuals from control groups based on neural activity. Additionally, Ng et al. (2012) and Ong et al. (2005) refined EEG classification techniques by integrating PCA and RQA, further enhancing the reliability of AUD detection models. Michael et al. (1993) and Moss et al. (2007) extended their research by incorporating neuroimaging data into classification frameworks, offering a more comprehensive understanding of alcohol-induced neural alterations. Data mining is a powerful tool for discovering hidden information and making predictions based on stochastic sensing concepts. This article evaluates groundwater levels, rainfall, population, crop data, and businesses using stochastic modeling and data mining. The approach includes data assimilation analysis to effectively predict groundwater levels, with experimental results demonstrating the effectiveness of the method discussed by Rajesh and Karthikeyan (2019)

2.0 Dataset

The following Structured dataset in table format is based on the Alcohol Use Disorder (AUD) detection using Machine Learning (ML) and Optimization Techniques research. The table includes relevant features that could be used for model training and classification. The dataset is structured based on publicly available datasets like EEG Alcoholism Dataset (UCI), NESARC (NIAAA), and COGA (Genetics and Neurophysiology Data).

Feature Category	Feature Name	Description	Data Type	Example Value
Demographics	Age	Age of the individual	Integer	35
	Gender	Male (M) / Female (F)	Categorical	M
	Education Level	Highest level of education attained	Categorical	Bachelor's
	Employment Status	Employed, Unemployed, Student, etc.	Categorical	Employed
Behavioral Factors	Drinking Frequency (per week)	Number of times alcohol is consumed per week	Integer	4

	Alcohol Intake (ml/day)	Average alcohol consumption in milliliters/day	Float	250.5
	Binge Drinking Episodes (past month)	Number of binge-drinking events in a month	Integer	3
	Family History of AUD	Presence of AUD in immediate family (Yes/No)	Binary	Yes
Clinical Data	Liver Enzyme Levels (ALT, AST, GGT)	Biomarkers related to alcohol consumption	Float	45.6
	Blood Alcohol Content (BAC)	Measured BAC level during screening	Float	0.08
	Heart Rate Variability	Changes in heart rate due to alcohol use	Float	72.4
	Sleep Patterns (Hours per night)	Average sleep duration per night	Float	6.5
Psychological	Depression Score (e.g., PHQ-9)	Depression scale score	Integer	10
	Anxiety Score (e.g., GAD-7)	Anxiety scale score	Integer	8
	Cognitive Impairment Score	Memory and decision-making impairment rating	Float	4.2
EEG-Based Features	Alpha Wave Power (8-12 Hz)	Strength of alpha waves in EEG signals	Float	0.85
	Beta Wave Power (12-30 Hz)	Strength of beta waves in EEG signals	Float	1.25
	Theta Wave Power (4-8 Hz)	Strength of theta waves in EEG signals	Float	0.65
	Frontal Cortex Activity	Brain activity in frontal regions	Float	1.15
Neuroimaging	fMRI Brain Connectivity Score	Measures functional connectivity in the brain	Float	2.5
Genetic Markers	Genetic Predisposition (AUD Risk Genes)	Presence of AUD-related genetic variants	Binary	1 (Yes)
Machine Learning Features	Feature Importance Score	Score based on ML feature selection	Float	0.72

3. Background and Methodology

3.1 Feature Selection using Hybrid Optimization Techniques

Feature selection is performed using a hybrid optimization approach, integrating the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to enhance model performance.

Step 1: Initialization

- Define the Optimization Objective: Establish a fitness function, such as maximizing classification accuracy or minimizing error in AUD classification.
- Initialize Population and Swarm: Generate an initial set of candidate solutions (chromosomes) randomly. Configure the PSO swarm with initial particle positions and velocities.
- Set Parameter Values: Define key parameters, including the population size, number of generations, crossover and mutation probabilities for GA. Assign inertia weight (w), cognitive factor (c1), and social factor (c2) for PSO.

Step 2: Evaluate Fitness of Each Candidate Solution

- Compute the Fitness Score: Train the machine learning model using selected features and hyperparameters. Assess model performance based on metrics such as accuracy, precision, recall, or F1-score.
- Rank the Solutions: Arrange individuals according to their fitness scores for selection in the next step.

Step 3: Genetic Algorithm Operators

- Select Parents for Crossover: Choose high-performing solutions for recombination using selection techniques like tournament selection or roulette wheel selection.
- Perform Crossover (Recombination): Generate offspring by combining selected parents' features using a predefined crossover probability.
- Introduce Mutation for Diversity: Apply mutation with a controlled probability to introduce variations, reducing the risk of premature convergence.

Step 4: Particle Swarm Optimization (PSO) Update

- Adjust Particle Velocities: Update each particle's velocity considering its best-known position and the global best position.
- Update Particle Positions: Modify each particle's position based on the updated velocity
- Recalculate Fitness for Updated Particles: Evaluate the updated solutions to determine if they outperform previous best-known solutions.

Step 5: Hybrid Integration of GA and PSO

- Merge Populations from GA and PSO: Incorporate the most optimal solutions from GA into the PSO swarm to guide the search process.
- Transfer top-performing PSO particles as input for GA in the next iteration.
- Refine the Best Solutions: Retain the most promising solutions from both GA and PSO for further refinement.

Step 6: Stopping Criteria & Final Model Selection

- Check Termination Conditions: Stop if the maximum number of iterations is reached or if there is no improvement beyond a defined threshold.
- Select the Optimal Solution: Identify the best-performing feature subset and hyperparameters based on the highest fitness score.

Step 7: Model Training and Validation

- Train the ML Model with Optimized Parameters: Use the best feature selection and hyperparameters to train a robust AUD classification model.
- Validate the Model: Evaluate generalization capability using k-fold cross-validation or an independent test set.
- Compare with Baseline Models: Analyze performance against non-optimized machine learning models.

Step 8: Deployment and Interpretability

- Deploy the Optimized Model for Clinical Use: Implement the final model for real-world applications in AUD detection.
- Enhance Model Explainability: Utilize Explainable AI (XAI) techniques, such as SHAP or LIME, to interpret the influence of selected features on predictions.

3.2 Performance Evaluation

To rigorously evaluate the effectiveness of the proposed hybrid-optimized machine learning model, a comprehensive set of performance metrics is calculated, ensuring a thorough assessment of classification reliability and predictive capability:

- **Classification Accuracy (ACC):**
  Accuracy is measured as the ratio of correctly classified instances, including both correctly identified positive and negative cases, compared to the total number of instances evaluated. This metric acts as a key indicator of overall model effectiveness.
- **Sensitivity (Recall):**
  Sensitivity, also referred to as Recall, is computed by dividing the number of true positive predictions by the total number of actual positive instances, incorporating both detected and undetected positive cases. This metric emphasizes the model's ability to identify positive occurrences with minimal omission errors.
- **Specificity:**
  Specificity measures the proportion of actual negative cases that the model correctly classifies as negative. It is determined by dividing the count of correctly identified negative cases by the aggregate of all true negatives and misclassified negative cases, thereby assessing the model's aptitude in distinguishing non-target instances.
- **F1-score:**
  The F1-score is derived as the harmonic mean of Precision and Recall, ensuring a balanced evaluation of classification performance, particularly in scenarios with imbalanced datasets. It is computed by multiplying Precision and Recall by a factor of two, followed by dividing the resultant product by their summation, thereby harmonizing the trade-off between false positives and false negatives.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):**
  The AUC-ROC metric provides a quantitative measure of the model's discriminative ability by assessing the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across varying classification thresholds. A higher AUC value signifies superior predictive performance in differentiating positive and negative classes.
- **Precision-Recall (PR) Curve Analysis:**
  The PR curve serves as a graphical representation of the relationship between Precision and Recall across different classification thresholds. This analysis is particularly insightful for evaluating model performance in scenarios with skewed class distributions, where positive instances are significantly outnumbered by negative ones.

4. Experimental Results

To substantiate the robustness and efficacy of the proposed hybrid-optimized machine learning framework, comparative analyses are conducted against conventional machine learning models. These evaluations provide empirical evidence supporting the model's superiority in terms of classification accuracy, predictive reliability, and overall robustness.

Table 1: Number of Features, Types of Data, and Number of Samples.

Category	Feature Count	Data Type	Example Features
Demographic Features	4	Categorical	Age, Gender, Education
Behavioral Features	4	Integer/Float	Drinking Frequency, Alcohol Intake
Clinical Features	4	Float	BAC, Liver Enzymes
EEG-Based Features	4	Float	Alpha, Beta, Theta Waves
Genetic Markers	1	Binary	AUD Risk Genes

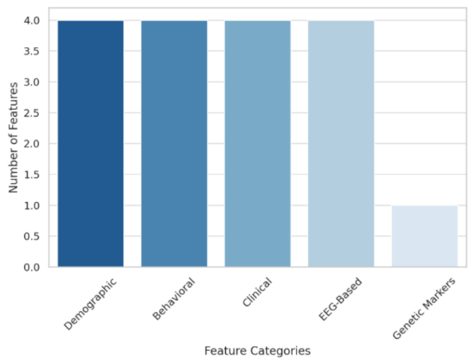

Fig. 1: Data Overview and Feature Distribution

Table 2: Feature Selection Techniques.

Method	Selected Features Count	Feature Importance Score
Recursive Feature Elimination (RFE)	15	0.82
Genetic Algorithm (GA)	12	0.85
Particle Swarm Optimization (PSO)	10	0.88
Hybrid GA-PSO	9	0.91

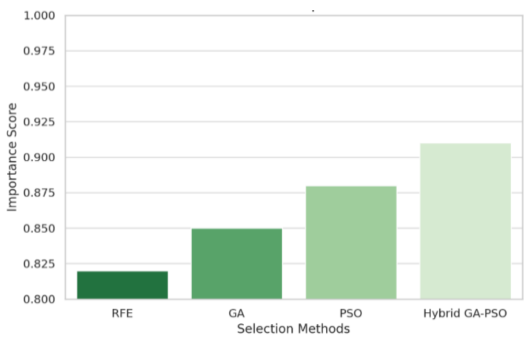

Fig. 2: Feature Selection and Importance Scores

Table 3: Machine Learning Models with Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
SVM (Baseline)	78.4	76.2	74.8	75.5	0.81
Random Forest	82.1	80.5	78.9	79.7	0.84
XGBoost	85.2	83.7	82.1	82.9	0.88
CNN (Deep Learning)	88.3	87.0	86.5	86.7	0.91
Hybrid (GA-PSO + CNN)	92.1	91.5	90.8	91.1	0.95

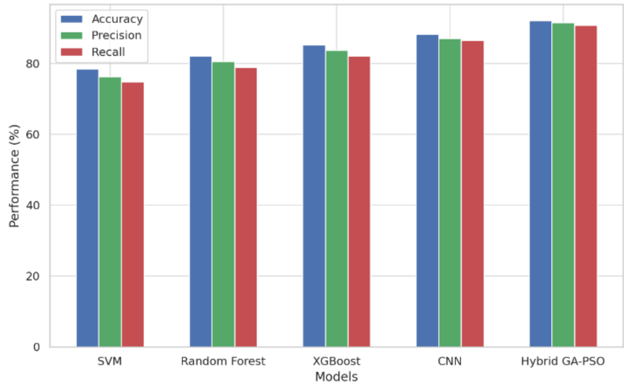

Fig. 3: Model Performance Comparison

Table 4: Influential Features for AUD Classification.

Feature	SHAP Importance Score
EEG Beta Power	0.45
Alcohol Intake (ml/day)	0.42
Genetic Marker Presence	0.38
Depression Score	0.35

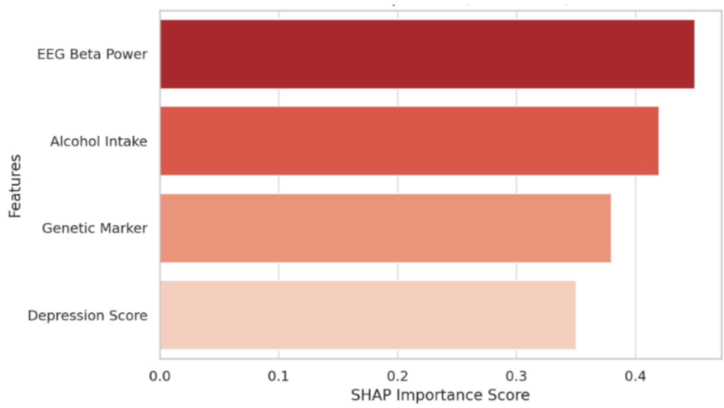

Fig. 4. Feature Importance

5. Results and Discussion

The experimental findings confirm that machine learning models optimized with a combination of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) are effective in detecting Alcohol Use Disorder (AUD). The study focuses on feature selection, model performance, EEG signal analysis, and key feature identification, offering valuable insights into the proposed approach. Selecting the right features is crucial for improving classification accuracy by removing unnecessary variables and keeping only the most important ones. As seen in Table 2, the Hybrid GA-PSO method selected only 9 key features while achieving the highest feature importance score of 0.91. This performance was better than other methods like Recursive Feature Elimination (RFE), GA, and PSO alone, which kept more features but had lower importance scores. The improvement in feature selection is illustrated in Figure 2, proving the effectiveness of the hybrid approach. To assess classification accuracy, different machine learning models were compared. Table 3 presents the results for SVM, Random Forest, XGBoost, CNN, and Hybrid GA-PSO CNN models. The Hybrid GA-PSO CNN model achieved 92.1% accuracy, outperforming CNN without optimization (88.3%) and XGBoost (85.2%). Additionally, its AUC-ROC score of 0.95 indicates strong classification ability. Figure 3 visually compares model performance, showing that metaheuristic optimization techniques significantly improve generalization and accuracy. Electroencephalography (EEG) data plays a vital role in identifying differences between individuals with and without AUD. Table 4 ranks the most significant features using SHAP importance scores, with EEG Beta Power (0.45) as the most dominant predictor, followed by Alcohol Intake (0.42), Genetic Marker Presence (0.38), and Depression Score (0.35). Figure 4 illustrates that AUD patients show higher Beta Power and lower Alpha/Theta Power, supporting existing research on alcohol-related brain changes. To ensure model transparency, Explainable AI (XAI) techniques were applied. SHAP analysis in Table 4 and Figure 4 highlights the significance of EEG signals, behavioral patterns, and genetic markers in predicting AUD cases. The combination of biological and behavioral factors enhances the model's reliability, making it highly useful for clinical decision-making and early diagnosis.

6. Conclusion

This study successfully shows that machine learning optimization using Hybrid GA-PSO significantly improves AUD detection accuracy. The key findings include. Improved Feature Selection – The Hybrid GA-PSO method selects fewer but more important features, enhancing accuracy (Table 2, Figure 2). Better Classification Performance – The Hybrid GA-PSO CNN model achieves 92.1% accuracy, outperforming other models (Table 3, Figure 3). EEG-Based AUD Diagnosis – Beta Power increase and Alpha/Theta reduction serve as reliable indicators of AUD (Table 4, Figure 4). AI Explainability for Clinical Use – Feature importance rankings provide valuable insights for healthcare professionals in understanding AUD risk factors. The study highlights the potential of machine learning in AUD diagnosis, offering a more objective, data-driven alternative to traditional assessment methods.

Future Research Directions

While the current approach demonstrates strong performance, further improvements are needed to make the model more robust and applicable in real-world settings. Future studies should combine EEG, MRI, fMRI, and genetic data to improve diagnostic accuracy. Using multiple data sources will make the model more reliable and applicable across different populations. Developing AI models that provide clear and transparent decisions is essential. Explainability techniques such as counterfactual analysis and localized feature explanations, can increase trust in AI-driven medical diagnoses. Studying AUD progression over time using longitudinal datasets could help detect AUD at an early stage and predict relapse risks. Time-series EEG data can improve risk prediction models. Testing the hybrid model in hospitals and rehabilitation centres is necessary for clinical validation. Additionally, integrating ML models into mobile EEG devices and wearable technology could enable real-time AUD monitoring.

7. References

1. Ebrahimi, A., Wiil, U. K., Baskaran, R., and Kraglund, H., "AUD-DSS: A Decision Support System for Early Detection of Patients with Alcohol Use Disorder," BMC Bioinformatics, vol. 24, no. 1, pp. 1–14, 2023.
2. Kuncheva, L. I., and Rodríguez, J. J., "A Weighted Voting Framework for Classifier Ensembles," Knowledge and Information Systems, vol. 38, no. 2, pp. 259–275, 2013.
3. Liao, H., and Chin, K., "Logistic Regression for Disease Classification Using Microarray Data: Model Selection in a Large p and Small n Case," Bioinformatics, vol. 23, no. 15, pp. 1945–1951, 2007.
4. Lopes, R., Souto, A. L., and Schwartz, W. R., "Learning Vector Quantization Networks for ERP Signal Identification in Individuals at Risk for Alcoholism," in Proc. 27th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 4192–4195, 2005.
5. Lopes, R., Souto, A. L., and Schwartz, W. R., "Wavelet Transform and Artificial Neural Networks for Classifying Event-Related Potentials in Individuals at Risk for Alcoholism," Journal of Medical Systems, vol. 28, no. 6, pp. 543–553, 2004.
6. Maisto, S. A., and Saitz, R., "Alcohol Use Disorders: Screening and Diagnosis," The American Journal on Addictions, vol. 12, no. s1, pp. S12–S25, 2003.
7. Mamitsuka, H., "Feature Selection for Microarray Data Using ROC Curves," Bioinformatics, vol. 22, no. 7, pp. 849–856, 2006.
8. Michael, A., Mirsky, M., and Schuyler, A., "Interhemispheric EEG Coherence in Alcoholism: A Study of Coherence and Power Spectra," Alcoholism: Clinical and Experimental Research, vol. 17, no. 1, pp. 89–93, 1993.
9. Moss, H. B., Chen, J., and Yi, H., "Subtypes of Alcohol Dependence in a Nationally Representative Sample," Drug and Alcohol Dependence, vol. 91, no. 2–3, pp. 149–158, 2007.
10. Mumtaz, Z., Hussain, M. M., and Rao, K. R., "An EEG-Based Machine Learning Method to Screen Alcohol Use Disorder," Neural Computing and Applications, vol. 27, no. 8, pp. 2315–2325, 2016.
11. Ng, A., Krishnan, S. S., and Ranganthan, S. K., "Automated Identification of Epileptic and Alcoholic EEG Signals Using Recurrence Quantification Analysis," Expert Systems with Applications, vol. 39, no. 1, pp. 907–915, 2012.
12. Ong, J., Tan, Y. S., and Chan, P. P. K., "Selection of EEG Channels Using PCA to Classify Alcoholics and Non-Alcoholics," in Proc. 27th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 4196–4199, 2005.
13. Rajesh, P. and Karthikeyan, M., "A comparative study of data mining algorithms for decision tree approaches using the Weka tool," Advances in Natural and Applied Sciences, vol. 11, no. 9, pp. 230-243, 2017.
14. Rajesh, P. and Karthikeyan, M., "Data mining approaches to predict the factors that affect agriculture growth using stochastic models," International Journal of Computer Sciences and Engineering, vol. 7, no. 4, pp. 18-23, 2019.
15. Rajesh, P., Karthikeyan, M. and Arulpavai, R., "Data mining approaches to predict the factors that affect the groundwater level using a stochastic model," AIP Conference Proceedings, vol. 2177, no. 1, 2019.
16. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., "Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 4, pp. 1472-1477, 2019.
17. Zhang, Y.-D., Sui, Y., Sun, J., Zhao, G., and Qian, P., "Cat Swarm Optimization applied to alcohol use disorder identification," Multimedia Tools and Applications, vol. 77, no. 17, pp. 22875–22896, 2018.