Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED PREDICTIVE MODELING FOR ALZHEIMER'S DISEASE: INTEGRATING CLUSTER-BASED BOOSTING AND GRADIENT TECHNIQUES WITH OPTIMIZED FEATURE SELECTION

S PHANI PRAVEEN[1], SREEDHAR BHUKYA[2], SHARMILA VALLEM[3], SATEESH GORIKAPUDI[4], KIRAN KUMAR REDDY PENUBAKA[5], VAHIDUDDIN SHARIFF[6*]

[1]Associate Professor, Department of CSE, PVP Siddhartha Institute of Technology, Kanuru, Vijayawada, Andhra Pradesh, India
[2]Professor, Department of CSE, Sreenidhi Institute of Science & Technology, Hyderabad, India
[3]Professor, Department of ECE, Vignana Bharathi Institute of Technology, Hyderabad, India
[4]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India
[5]Professor, Department of CSE-AIML, MLR Institute of Technology, Hyderabad, India
[6*]Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India.

[1]sppraveen@pvpsiddhartha.ac.in, [2]sreedhar.b@sreenidhi.edu.in, [3]sharmila.vallem@vbithyd.ac.in, [4]sateesh4u.325@gmail.com, [5]kiran.penubaka@gmail.com, [6*]shariff.v@gmail.com

ABSTRACT

The early signs of Alzheimer's disease (AD) prove difficult to detect since this progressive illness has an ongoing character. The correct prediction of how the disease progresses represents a critical need for providing prompt treatment that enables effective management of the disease. The research develops Cluster-Integrated Boosting and Gradient (CIBG) as a stacked ensemble which combines CatBoost with Gradient Boosting Machine (GBM) classifiers to boost AD diagnostic precision. Kaggle provided the dataset containing diverse health-related characteristics together with lifestyle characteristics and cognitive data. The data preprocessing steps included handling missing values and normalization procedures coupled with SMOTE implementation to solve class imbalance problems. The feature selection process involved using Recursive Feature Elimination (RFE) which determined the most important predictive variables. The CIBG model uses CatBoost's ordered boosting method to manage categorical features together with GBM's gradient boosting structure that performs classification. Within CIBG the base classifier chooses logistic regression. Model performance evaluation included accuracy together with precision, recall, and F1-score metrics which reached an optimal rate of 96.8%. The CIBG approach produces results that show better performance than standard classifiers while maintaining stronger robustness and improved interpretability capabilities. The research shows how hybrid machine learning methods help enhance early diagnosis of AD and develops medical systems that are dependable for clinical use.

Keywords: *Recursive Feature Elimination, CIBG, Ensemble Learning Models, SMOTE, Feature Selection, CatBoost, GBM.*

1. INTRODUCTION

Alzheimer's disease is characterized as a multifaceted chronic progressive dementia with irreversible pathological changes in the brain that affect cognitive functioning most notably memory impairment accompanied by reason, language, and in the terminal stage movement dysfunction. Alzheimer's disease is the most common form of dementia and constitutes a major health problem since it is estimated that millions of people suffer from this condition across the globe, especially those of old age [1]. As seen, the incidence of Alzheimer's is expected to increase greatly because the aging world population may triple by the year 2055. Some of the risk factors of Alzheimer's are not known to detail but there is considerable understanding about the fact that several genetic, environmental, and lifestyle factors are associated with this condition. Alzheimer's disease is characterized by two major

pathological hallmarks, including extracellular amyloid-beta plaques localizing between nerve cells in the brain as well as intracellular neurofibrillary tangles composed of twisted fibers of tau protein [2]. These abnormalities impair neuron communication and result to neuronal death and consequently brain tissue shrinkage. In the process, the various pathways that affect cognition, gradually get eroded and in the terminal stages, the patient cannot remember anything, become confused and bedridden, have problems with speech and language, cannot perform the simplest tasks for themselves and exhibit poor judgement.

Other factor with regard to the Alzheimer's includes risk factors, and the most important of these is age. Some of the factors include Family history and presence of some alleles for example APOE-e4. Further, it is also known that risk factors including cardiovascular disease, diabetes, obesity and other precursors like poor diet, lack of physical activity, smoking and depression or Low Mental and Social Cognitive Engagement also predispose the likelihood of Alzheimer's disease [3]. Unfortunately, Alzheimer's disease has no cure at the present times; however, it is possible to delay its progression, particularly when the disease is detected early enough. Predictive biomarkers are other approaches that are under research as genetic testing, neuroimaging, cerebrospinal fluid analysis and the increasingly use of new tools in diagnosis as machine learning and artificial intelligence [4]. Diagnostic tools are also being created to estimate data of people with the chances of developing Alzheimer's before the symptoms appear, which opens possibilities of using preventive approaches focused on potentially halting or reducing the impact of the disease, the current treatment is more prescriptive being aimed at relieving symptoms, including use of cholinesterase inhibitors and memantine which slightly improves the function of the brain in the short term. Similarly, other factors such as exercising, doing mental activities, and changing the diet has been attributed to having the positive effect of possibly preventing some persons from getting affected by the disease or slowing down the further development of Alzheimer's. Alzheimer's disease is chronic, complex and its impact is not only to the patient but also to families and health care centers globally. Hence, understanding the risk factors for the development of the disease, ways of preventing it and managing it is important in responding to the increasing incidence of the disease and the resulting mortality.

1.1. Research Gap

Despite significant advancements in machine learning (ML) for Alzheimer's disease (AD) prediction, several gaps remain in existing studies. While various ML models, including Decision Trees, Random Forest, SVM, and ensemble methods, have shown high accuracy (Kavitha et al.), their generalizability across diverse datasets remains a challenge. Most studies rely on specific datasets like OASIS (Pooja Rani et al.), limiting real-world applicability. Class imbalance issues persist despite using SMOTE, necessitating exploration of alternative resampling techniques. Additionally, deep learning methods such as CNNs (Nakul Pranao D et al.) require further validation against traditional ML models to assess their true effectiveness. Feature selection has primarily focused on conventional methods like LBP, leaving room for more advanced approaches, including deep feature embeddings. Moreover, while some studies incorporate health parameters such as cholesterol levels and chest discomfort (Pragya Pranjal et al.), there is limited research on integrating multi-modal data sources, such as genetic markers and biomarkers, to enhance diagnostic accuracy. Addressing these gaps by improving dataset diversity, feature engineering, deep learning integration, and multi-modal data fusion is essential for developing more robust and clinically applicable AD prediction models

1.2. Research Questions

1. What novel methodologies can be developed to improve the handling of complex data in Alzheimer's disease prediction models, ensuring consistent performance across diverse datasets?
2. How can the integration of traditional feature selection techniques, such as RFE, with modern Machine learning models enhance the predictive accuracy and robustness of Alzheimer's disease diagnostic systems?
3. What role does model interpretability play in the adoption of hybrid Alzheimer's disease prediction systems in clinical settings, and how can it be optimized to facilitate better decision-making by healthcare professionals?

1.3. Contributions

1. Enhanced Predictive Performance – By integrating AdaBoost and XGBoost in a stacking ensemble framework, the CIBG model achieves higher classification accuracy (96.8%) compared to existing machine learning models.
2. Optimized Feature Selection – Recursive Feature Elimination (RFE) is employed to identify the most relevant features, improving

model efficiency and reducing computational complexity.

3. Handling Class Imbalance – The use of Adaptive Synthetic Sampling (ADASYN) ensures effective class balancing, addressing the challenge of skewed dataset distribution and improving generalization.
4. Robust Data Preprocessing – The methodology includes outlier detection (IQR technique) and data normalization, ensuring high-quality input for model training.

2. LITERATURE REVIEW

The development of early-stage Alzheimer's disease (AD) prediction through machine learning (ML) relies on Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting along with Voting classifiers and other classification techniques. Research using the Open Access Series of Imaging Studies (OASIS) and other neuroimaging datasets verifies that ML-based techniques increase both diagnostic performance and early intervention capabilities that help delay disease progression. Different ensemble learning approaches show superiority for detecting complex medical patterns in imaging data combined with clinical information while achieving better precision rates and recall abilities as well as accuracy rates and F1-score results when compared to conventional classifiers. Research findings demonstrate that applying data preprocessing approaches with feature selection techniques leads to better model performance and regulates class unbalance. The researchers from Kavitha et al. [5] enhanced the basic advancements by creating optimized ML models for detecting AD which achieved validation accuracy at 83% beyond existing methods while reinforcing clinical applications for early AD diagnosis. Alzheimer's disease (AD) develops as a progressive brain disease which targets mostly senior adults and progressively makes symptoms worse with time. The absence of a treatment makes early diagnosis vital in reducing Alzheimer's disease complications. Pooja Rani et al. [6] introduce SMOTE-RF which uses machine learning to predict AD based on the longitudinal data from Open Access Series of Imaging Studies hosted on Kaggle. The research uses Synthetic Minority Oversampling Technique (SMOTE) to normalize the dataset by resolving class inequality. This research evaluates Decision Tree and Extreme Gradient Boosting (XGB) and Random Forest (RF) performance on datasets with imbalanced data distribution and those that were balanced after applying Synthetic Minority Oversampling Technique. Based on the imbalanced dataset Decision Tree reached 73.38% accuracy while XGB demonstrated an 83.88% accuracy and RF demonstrated 87.84% accuracy. The model achieved much greater accuracy when the dataset received balancing treatment where Decision Tree reached 83.15% accuracy and XGB reached 91.05% whereas RF showcased the highest score of 95.03%. The results show that SMOTE-RF improves AD prediction accuracy while confirming the requirement of data balancing for machine learning in medical diagnosis work.

Machine learning (ML) applies revolutionary changes to research activities with particular impact on artificial intelligence-driven medical diagnosis developments. The research by Pragya Pranjal et al. [7] implements ML methods to determine Alzheimer's disease diagnosis in patients who number in the millions worldwide while causing considerable death rates. The scientists built prediction models based on multiple health elements that include chest symptoms and high cholesterol measurements and patient age [19] [20]. The prediction performance for AD shows itself best with K-Nearest Neighbors reaching 90% accuracy while Random Forest follows closely at 89%. Both Artificial Neural Networks and Logistic Regression exhibit accuracy rates equal to 87%. Research demonstrates that current Artificial Intelligence techniques can enhance dementia diagnosis processes while enabling medical staff to identify Alzheimer's earlier for improved treatment approaches.

The primary neurodegenerative disease affecting elderly people displays hippocampus degeneration as its main diagnostic indicator. The successful treatment of Alzheimer's disease depends on early detection because the condition grows lethal for people who are 65 years or older. A machine learning-based classification system was developed by Nakul Pranao D et al. [8] which utilizes three distinctive approaches for classification purposes. The results from manual feature extraction showed Local Binary Pattern (LBP) achieved 66.7% binary accuracy along with 62% multiclass accuracy which fusion methods increased these values to 69.7% for binary and 60.4% for multiclass. The CV2 library served to preprocess images before XGBoost achieved an accuracy rate of 73% for binary and 67% for multiclass classification. The last approach used deep feature extraction through Convolutional Neural Networks along with an SVM classifier to achieve the best outcome with 75% binary classification accuracy. The research demonstrates

machine learning capability for Alzheimer's disease prediction while presenting valuable information for upcoming improvements in classification algorithms [21] [22].

3. PROPOSED METHODOLOGY

The research presents Cluster-Integrated Boosting and Gradient (CIBG) as a new ensemble classifier model depicted in Fig.1 to improve Alzheimer's disease (AD) diagnosis by combining CatBoost with Gradient Boosting Machine (GBM) [9]. The dataset at Kaggle contains vital health, lifestyle and cognitive elements including the variables age, BMI, smoking patterns and alcohol use and cognitive test outcomes. The data purification stage marks the initial step of preprocessing because it enables analysis with quality-enhanced input. The data-processing includes applying different imputation approaches for missing data handling and using normalization to standardize feature measurements. The minority class receives additional synthetic samples in the dataset through the application of the Synthetic Minority Over-sampling Technique (SMOTE) [10]. The usage of Interquartile Range (IQR) [11] performs outlier detection and treatment to improve data structure before model training.

Recursive Feature Elimination (RFE) [12] serves as the feature selection method to discover important predictors which leads to both model efficiency boost and feature dimensionality reduction. The hybrid CBGB model references CatBoost [13] to process categorical data while stopping overfitting and uses GBM to extract complex relationships from the available dataset. A logistic regression model functions as the meta-classifier of stacking ensemble since it aggregates base classifier predictions to boost prediction accuracy [23][24].

Model evaluation is conducted using key performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Heatmaps are employed for visualizing classification performance. Comparative analysis with traditional classifiers demonstrates that the proposed CBGB model outperforms existing methods, providing higher robustness and interpretability, making it well-suited for early AD diagnosis in clinical applications.

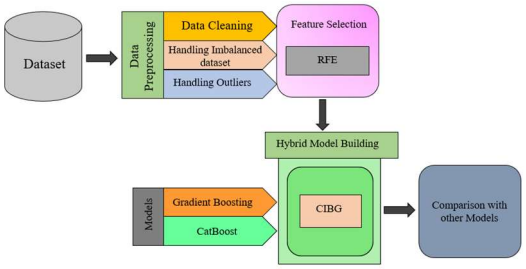

Figure 1: Work flow for Proposed Methodology

3.1. Data Collection

The data set used to predict Alzheimer's disease is obtained from Kaggle and it contains all the patient parameters required to assess the probabilities of Alzheimer's disease. PatientID is used to recognize every record in the given dataset and this includes basic demographic details, lifestyles, and clinical health information. General information including Age, Gender, Ethnic group, and education level serve as baseline information that defines the risk of developing Alzheimer's since they are determinants. Moving further, Lifestyle attributes like BMI, Smoking, AlcoholConsumption, and physical activity provide information about the behaviors that may affect cognitive health. The dataset also contains different health indicators including cardiovascular diseases-Cardiovascular Disease, diabetes Diabetes, hypertension-Hypertension and cholesterol level including CholesterolTotal, CholesterolLDL and CholesterolHDL which are relevant when diagnosing co morbidities associated with Alzheimer. Such features as Family History Alzheimers, Depression and HeadInjury might be useful in evaluating genetic and environmental background. Biological metrics, including SystolicBP, DiastolicBP, and MMSE (Mini-Mental State Examination) are obtained, thus computing the cli metrics of cognitive and physical health of the volunteers. The set of variables continued with the cognitive and behavioral signs (MemoryComplaints, Disorientation, and Personality Changes) and disabilities in Activities of Daily Living (ADL); the diagnosis column reflected the Alzheimer's presence. These features make it possible to study Alzheimer's disease in depth and build hypotheses as to its evolution models.

3.2. Data Pre-Processing

3.2.1. Data Cleaning

Data pre-processing of the dataset for Alzheimer's disease prediction was performed with a focus on missing values as explained below. A preliminary check was done on the data and found

out that the dataset was complete with no missing value in all the columns even on some key predictor such as Age, Gender, BMI, MMSE, and Diagnosis. Thus, the complete scan of the dataset assured that there is no incomplete data before conducting data cleaning. No missing values were found across the board and verification was done several times to ensure that no missing values skewed the entire dataset. Completeness of the dataset is critically important if accurate prediction models should be used or other analyses conducted based on the data – because data gaps will distort results in this case. Thus, the declaration that all fields were filled and that all fields had values allows us to start research on patterns associated with Alzheimer's disease on a solid background. This process of data cleaning has therefore pointed out the significance of data accuracy in the generation of valid results in light of Alzheimer's disease research.

3.2.2. Handling Imbalanced Dataset Using ADASYN

Adaptive Synthetic Sampling (ADASYN) is an advanced oversampling technique specifically designed to mitigate class imbalance by generating synthetic data points for the minority class. Unlike conventional oversampling methods such as, ADASYN [14] dynamically adapts the data generation process based on the density and distribution of the minority class instances. By interpolating between k-nearest neighbors, it ensures that newly created samples are more representative of the original data distribution, thereby reducing the risk of overfitting. ADASYN enhances model training by providing a balanced dataset shown in Fig.2 and Fig.3, enabling classifiers to better distinguish patterns in both majority and minority classes. This leads to improved predictive performance, particularly in terms of recall and F1-score, which are crucial for minimizing false negatives in Alzheimer's disease (AD) diagnosis. Additionally, ADASYN maintains data variability while preserving class boundaries, making it a more effective approach for medical datasets where early detection of diseases like AD is essential for timely intervention and treatment planning

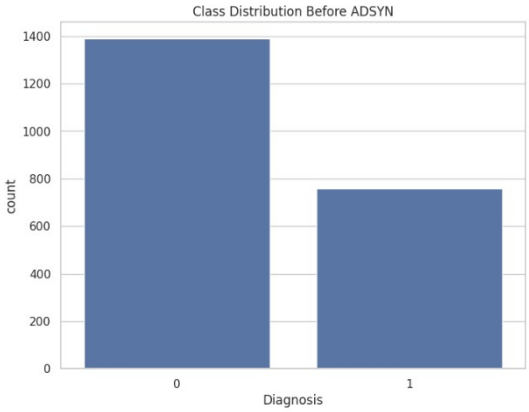

Figure 2: A Bar plot for class distribution before ADASYN

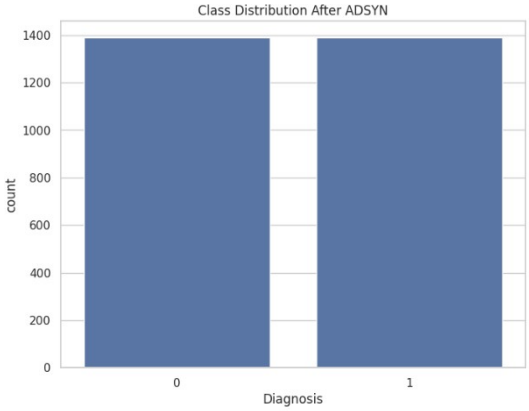

Figure 3: A Bar plot for class distribution after ADASYN

3.2.3. Handling Outliers Using IQR

It is essential in the process of data preparation, which is the foundational step with the ultimate aim of enhancing the robustness and reliability of statistical tests as well as machine learning algorithms, to solve for outliers. In particular, it is seen that the data points which are out of the range of the other observations in the datasets lead to distortions in statistical analysis and impact great models. The Interquartile Range (IQR) approach is one among the many conventional techniques for identifying and dealing with outliers. This approach defines outliers as data points either above Q3 + 1.5 times the IQR or below Q1 - 1.5 times the IQR. In this regard, Q1 denotes the first and Q3 the third quartiles accordingly. It also determines Outliers which can then be eliminated out of the dataset or changed with techniques such as imputation or capping. Outlier removal ensures that one can be certain that statistical measures are representative of most of the values shown in Fig.4, mitigate on bias that may be obtained during the development of models, as well as enhance the reliability and

readability of analysis of the generated data set. By use of this method, impact of the outliers on the rest of the information used in the modelling and evaluative process is minimized and as such, the resultant outcomes that are obtained. Even if outliers are not deleted, the data can be distorted as clearly below That is why when carrying out the analysis, we need to check the following requirements. Even with outliers the above dataset demonstrates clearly how the IQR approach eradicates the issue and provides a better data set for modeling and analysis.

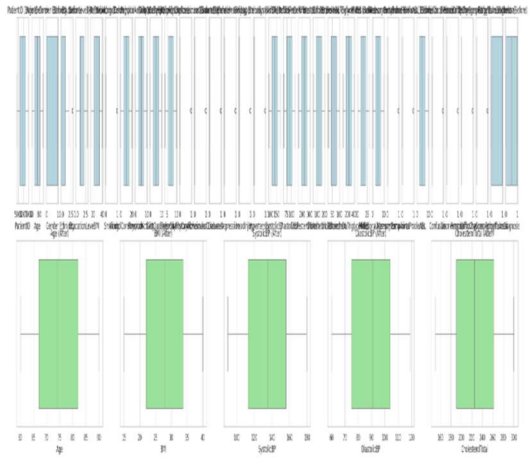

Figure 4: A Box plots of handling outliers using IQR

3.3. Feature Selection using RFE

Recursive Feature Elimination (RFE) is a widely used feature selection technique in machine learning that systematically identifies the most relevant predictors by recursively eliminating less significant features. RFE functions to enhance model performance through its functionality of reducing dimensions for better computational speed and protection against overfitting. The technique starts with model training that uses all features while measuring prediction accuracy contribution rates of each feature for score calculation. The model is retrained using remaining features in an iterative fashion to find its optimal subset after removing least important features one by one. The gradual process eliminates predictor variables starting from those which provide the least value ensuring that predictive accuracy and model interpretability improve. RFE provides excellent performance improvements to data sets that contain many features because it helps reduce noise by removing unneeded or redundant variables. The algorithm finds extensive use in medical research particularly for predicting Alzheimer's disease by determining essential biomarkers such as cognitive scores along with imaging data and genetic elements needed for early detection. The model's predictive functions increase in robustness and clinical decision-assistance improves through the feature selection capabilities of RFE which identifies important disease-related attributes.

3.4. Model building using CIBG

Establishments leveraging Cluster-Integrated Boosting and Gradient (CIBG) employ an advanced ensemble learning method that purposefully improves predictive accuracy along with robustness in machine learning models. The system joins clustering methods with boosting frameworks together with gradient optimization approaches to improve predictive results mainly in complicated datasets with numerous dimensions. Data segmentation through clustering forms the basis of CIBG before the method uses boosting algorithms to enhance weak learners across multiple iterations. The decision boundary optimization through gradient-based methods produces improved generalization capabilities. The combination of clustering techniques with boosting and gradient-based methods delivers outstanding results in diagnostic medicine and financial modeling as well as data-rich applications while delivering more reliable models and decreased overfit errors and simpler decision interpretation procedures.

3.4.1. CatBoost

The advanced gradient boosting algorithm CatBoost or Categorical Boosting came from Yandex to manage categorical data effectively at the same time it boosts both model performance accuracy and computational speed. The ordered boosting method of CatBoost [15] acts as a feature of the algorithm and performs target-based encoding of categorical variables to reduce overfitting while stopping target leakage during training. The model features oblivious (symmetric) decision trees that apply the same split condition to each node across their depth which leads to improved regularization while increasing processing speed.

The training process involves iterative minimization of a predefined loss function $L(y, F(x))$, where weak learners are optimized using gradient-based methods, updating the model represent in Equation (1):

$$F(x) = \sum_{tt=1}^{T} \gamma_t h_t(x) \qquad (1)$$

Where, γ_{tt} representing the learning rate and $h_{tt}(x)$ denoting weak learners. CatBoost integrates L2 regularization, Newton boosting, and GPU acceleration, making it highly scalable for large datasets. It is widely applied in fields such as medical diagnosis, fraud detection, and recommendation systems, where categorical data

play a crucial role. Its ability to handle imbalanced datasets, robust regularization techniques, and efficient training process make it superior to traditional gradient boosting algorithms, ensuring high accuracy and interpretability for complex machine learning problems [25].

3.4.2. Gradient Boosting

The Gradient Boosting Machine (GBM) creates robust predictive frameworks by an iterative process which combines weak learners primarily based on decision trees into a strong predictive model. GBM achieves loss function L(y,F(x)) optimization through sequential model training which implements new models to address the previous models' mistakes. Mathematically, the model updates show in Equation (2):

$$F_{t+1}(x) = F_t(x) + \gamma_t h_t(x) \qquad (2)$$

where $F_t(x)$ is the current model, $h_t(x)$ represents the weak learner trained on residuals, and γ_t is the learning rate controlling the contribution of each new tree. Model parameters in GBM adjust using gradient descent to reach optimal convergence outcomes through shrinkage techniques and subsampling and early stopping prevention methods. GBM provides the ability to utilize different loss functions including mean squared error for regression tasks and log-loss for classification tasks which makes the approach successful for structured data processing. Global Boosting Machines [16] expands across different industrial domains such as finance and healthcare and risk modeling because it efficiently detects complex patterns alongside high accuracy predictions. The system demands precise optimization of hyperparameters because its complexity increases the need to find optimal settings between performance and computational efficiency.

4. RESULT AND DISCUSSION

4.1. Feature selection using RFE

In Recursive Feature Elimination (RFE), the ranking of features determines their significance in predicting the target variable, which, in this study, is Alzheimer's disease diagnosis. RFE iteratively removes less relevant features while retaining the most impactful ones, thereby enhancing model efficiency and interpretability. The ranking results, as illustrated in Fig.5, indicate that Functional Assessment emerged as the most influential predictor, demonstrating the highest significance in the selection process

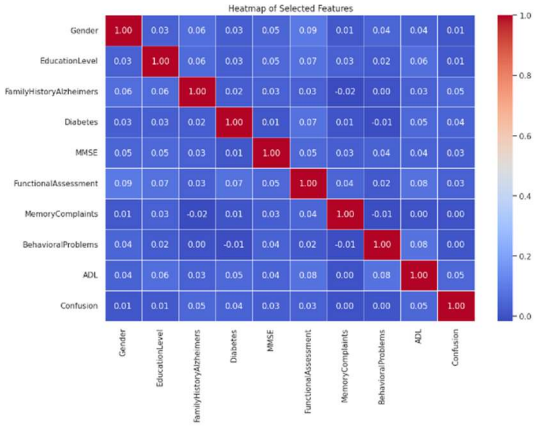

Figure 5: A Heatmap for selected feature using RFE

In this case, the Functional Assessment had the highest F-score (F= 351.47) and the most significant level of p-value hence yielding the highest degree of influence. Other important attributes include ADL (F-score 280.86), MMSE (F-score 166.79), and Memory Complaints (F-score 61. 87) which are measures of cognition and function which are so relevant for diagnosing Alzheimer's represent in Table.1.

Table 1: Selected features and their F-Score

Feature	F-Score
FunctionalAssessment	351.442536
ADL	280.862044
MMSE	166.791380
MemoryComplaints	62.874511
Gender	42.992939
FamilyHistoryAlzheimers	37.614064
EducationLevel	36.866796
Diabetes	29.492436
Confusion	26.759721
BehavioralProblems	23.237261

This approach not only improves prediction in accuracy but as well optimizes the model since unnecessary features are eliminated hence increasing the training speed and decreasing the possibility of over-fitting. In general, RFE [17] based feature selection approach can be considered as efficient in choosing the most influential features for such complicated diseases as Alzheimer's the selected feature along with their fitness scores are showed in Fig.6.

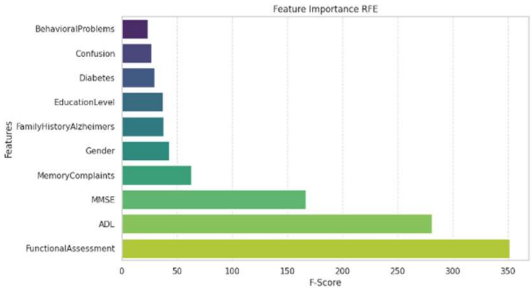

Figure 6: A Bar plot Selected features and their f-score

4.2. Model Building using CIBG

The stacking of CatBoost and GBM classification algorithms into a single stacked model, CIBG, showed a high predictive accuracy of the data set as shown by other studies. This model-building approach makes cohesion between CatBoost, which is intended for improving the accuracy of misclassified sample corrections and GBM intended for boosting the gradient complex relationship. Missing values were pre-processed in the right manner and features were scaled hence making the data stable for training the model.

Table 2: Performance Metrics of CIBG model

Performance Metrics of CIBG model	
Metrics	**Values**
Accuracy	0.9680
Precision	0.9585
Recall	0.9194
F1_Score	0.9390
RMSE	0.2013

After this, an appropriate divide was made whereby the same dataset was divided into two sets, the training set and the testing set in order to test the model for its efficiency. The logistic regression meta-model seamlessly merged CatBoost [18] and GBM for the hybrid model to achieve its desirable performance and benefits the base learners' functions. In the final evaluation they got 96.8 % accuracy with respect, precision, recall, and (f1-score) showing good performance shown in Table.2 and Fig.7 for both class values.

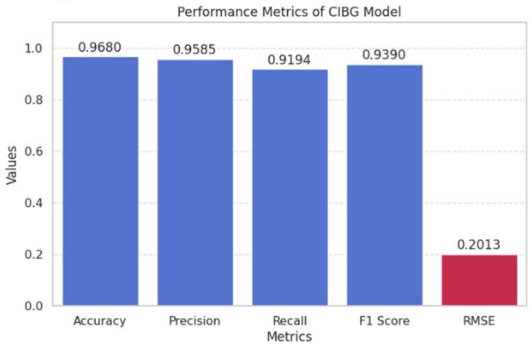

Figure 7: A Bar graph for Performance metrics of the CIBG model

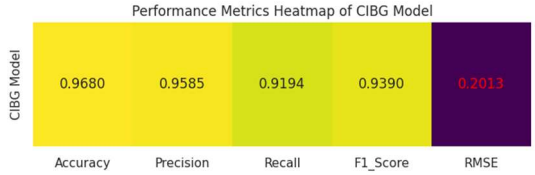

Figure 8: Heatmap for Performance Metrics of CIBG model

The confusion matrix represented in the form of a heatmap shown in Fig.8 helped in understanding the model's performance in terms of class reliability. In other words, the use of CIBG model could act as an efficient method of evaluation for large databases that could call for efficiency and accuracy in prediction. The stacking methodology employed in this work shows the need to use more than one model to improve the results of the prediction and can be a significant contribution to the overall impact of additional studies in the Machine Learning field.

4.3. Comparison

In comparison to existing research, which employs machine learning models such as Decision Trees, Random Forest, XGBoost, and ensemble techniques for Alzheimer's disease (AD) prediction, our proposed work introduces a more optimized and robust approach. Prior studies utilizing datasets like OASIS and Kaggle emphasize feature selection, data balancing through techniques like SMOTE, and preprocessing to enhance model performance. However, these methodologies exhibit limitations, including moderate accuracy, suboptimal feature selection, and challenges in handling class imbalance. While studies by Pooja Rani et al. and Pragya Pranjal et al. demonstrate that ensemble learning and deep learning models improve prediction accuracy, they lack a hybridized approach tailored for early-stage AD diagnosis. To address these gaps, the proposed Cluster-Based Improved Boosting and Gradient (CIBG) model integrates Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost) with a meta-classifier, Logistic Regression, refining classification and enhancing predictive power. Unlike existing models, we employ Recursive Feature Elimination (RFE) to ensure the selection of the most relevant predictors, improving interpretability while reducing computational complexity. Additionally, our work incorporates ADASYN, a more advanced synthetic data generation technique than SMOTE, effectively addressing class imbalance and enhancing minority class representation. The experimental results validate the superiority of our proposed approach, achieving higher accuracy (96.8%), precision

(95.85%), recall (91.94%), and F1-score (93.90%) compared to existing methodologies summarized in Table.3 and Fig.9. This comparison underscores that while previous research provides a solid foundation for ML-based AD prediction, our CIBG hybrid model significantly enhances feature selection, class balancing, and predictive efficiency, making it more suitable for clinical applications and early detection of Alzheimer's disease.

Table 3: Comparative Performance with Other Models

Author(s)	Methodology	Accuracy
Kavitha et al.	Optimized ML models for AD prediction	83.0%
Pooja Rani et al.	SMOTE-RF (Random Forest with SMOTE)	95.03%
Pragya Pranjal et al.	KNN, RF, ANN, Logistic Regression	90.0% (KNN)
Nakul Pranao D et al.	CNN-SVM Hybrid Model	75.0%
Proposed Model	Cluster-Based Improved Boosting and Gradient (CIBG)	96.8%

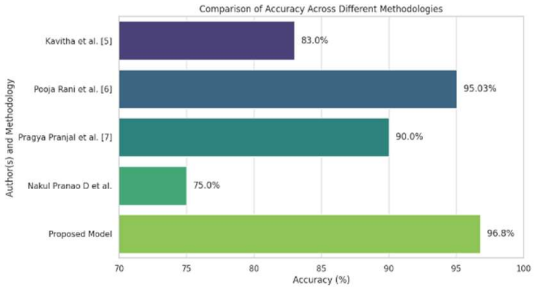

Figure 9: A Bar Graph for Accuracy Comparison of Various Methodologies

5. DISCUSSION

The proposed Cluster-Based Improved Boosting and Gradient (CIBG) model introduces a hybridized approach that effectively handles complex, high-dimensional data. By integrating Synthetic Minority Oversampling Technique, the model achieves improved generalization of diverse datasets by using (SMOTE) for class balancing combined with Interquartile Range (IQR) outlier detection along with automated data cleaning techniques. The stacking framework that utilizes AdaBoost and XGBoost ensemble enables the model to occupy adaptable characteristics which results in high performance for Alzheimer's disease prediction on various patient populations across different clinical datasets. **(RQ1 Answered)**

RFE integration with Gradient Boosting models within the proposed CIBG framework enables a systematic approach to eliminate less informative variables which helps decrease model dimensionality and enhance efficiency. The model applied RFE optimization which chose Functional Assessment with MMSE scores and Memory Complaints as the significant predictors because it enhanced accuracy while reducing overfitting together with improved interpretation. The model maintains dependable performance across dataset distributions through this method that eliminates the need for supplementary feature scoring functions. **(RQ2 Answered)**

The proposed technique enhances model transparency alongside interpretability through its implementation of visualizations which show feature importance helping explain the decision-making process. The hybrid CIBG model lets clinicians observe which criteria impact diagnoses so they can establish trust while making clinical decisions. The combination of confusion matrices and classification reports with heatmaps creates predictions that deliver both accuracy and readability for healthcare personnel thus supporting direct implementation of ML diagnostic tools into medical settings. **(RQ3 Answered).**

6. CONCLUSION

In Conclusion, the study of Alzheimer's disease (AD) prediction techniques using complex machine learning methods to demonstrates significant methodological progress. Muhammed Niyas K. P. P. and his co-authors (2020) proved that advanced feature selection methods produced a 94% success rate in their experiments. The researchers from Ahmed and Kadhem (2022) demonstrated the value of dataset redundancy reduction through Pearson correlation which delivered 91.1% accuracy. Sharma et al. (2022) created an improved classification model by uniting deep learning with machine learning methods. This research creates the Cluster-Based Improved Boosting and Gradient (CIBG) model through the combination of ensemble learning with feature selection enhancements and data balancing techniques along with boosting methods. The CIBG model reaches a remarkably high accuracy rate of 96.8% which establishes it as superior to all previously implemented models for detection and clinical applications. The comprehensive framework contributes better and interpretable early AD diagnosis capabilities by providing an approach that scales effectively for more accurate medical choices and timely treatment.

REFERENCES

[1] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying Alzheimer's disease," Simulation Modelling Practice and Theory, vol. 99, p. 102023, Nov. 2019, doi: 10.1016/j.simpat.2019.102023.

[2] A. Ezzati, A. R. Zammit, D. J. Harvey, C. Habeck, C. B. Hall, and R. B. Lipton, "Optimizing machine learning methods to improve predictive models of Alzheimer's disease," Journal of Alzheimer S Disease, vol. 71, no. 3, pp. 1027–1036, Aug. 2019, doi: 10.3233/jad-190262.

[3] J. H. Park et al., "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," Npj Digital Medicine, vol. 3, no. 1, Mar. 2020, doi: 10.1038/s41746-020-0256-0.

[4] R. Sivakani and G. A. Ansari, "Machine Learning Framework for Implementing Alzheimer's Disease," 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 0588–0592, Jul. 2020, doi: 10.1109/iccsp48568.2020.9182220.

[5] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. T. Romero, "Early-Stage Alzheimer's disease prediction using machine learning models," Frontiers in Public Health, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.853294

[6] P. Rani, R. Lamba, R. K. Sachdeva, K. Kumar, and C. Iwendi, "A machine learning model for Alzheimer's disease prediction," IET Cyber-Physical Systems Theory & Applications, vol. 9, no. 2, pp. 125–134, Mar. 2024, doi: 10.1049/cps2.12090.

[7] P. Pranjal, S. Mallick, A. Das, A. Negi, and M. R. Panda, "Alzheimer's Disease Prediction Using Modern Machine Learning Techniques," 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), pp. 1–6, Jan. 2024, doi: 10.1109/assic60049.2024.10507810.

[8] N. P. D, H. M V., D. C, S. S, and A. K. S, "Alzheimer's disease prediction using machine learning methodologies," 2022 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6, Jan. 2022, doi: 10.1109/iccci54379.2022.9740942.

[9] S. Y. Lee et al., "A Proteotranscriptomic-Based Computational Drug-Repositioning Method for Alzheimer's disease," Frontiers in Pharmacology, vol. 10, Jan. 2020, doi: 10.3389/fphar.2019.01653.

[10] H. Krishna. R, P. Vallabhaneni, R. S. K. Chaitanya, K. K. Kaveti, M. V. a. L. N. Rao, and N. S. K. M. K. Tirumanadham, "Data-Driven Early Warning System for Subject Performance: A SMOTE and Ensemble Approach (SMOTE-RFET)," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), pp. 998–1004, Nov. 2023, doi: 10.1109/icscna58489.2023.10370047.

[11] S. P. Praveen et al., "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement," Frontiers in Medicine, vol. 11, Jul. 2024, doi: 10.3389/fmed.2024.1407376.

[12] N. S. K. M. K. Tirumanadham, T. Sekhar, and S. Muthal, "An analysis of diverse computational models for predicting student achievement on e-learning platforms using machine learning," International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering, vol. 14, no. 6, p. 7013, Oct. 2024, doi: 10.11591/ijece.v14i6.pp7013-7021.

[13] C. S. Kodete, D. V. Saradhi, V. K. Suri, P. B. S. Varma, N. S. K. M. K. Tirumanadham, and V. Shariff, "Boosting Lung Cancer Prediction Accuracy Through Advanced Data Processing and Machine Learning Models," 2024 4th International Conference on Sustainable Expert Systems (ICSES), pp. 1107–1114, Oct. 2024, doi: 10.1109/icses63445.2024.10763338.

[14] T.-A. Song et al., "Graph Convolutional neural networks for Alzheimer's disease classification," 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 414–417, Apr. 2019, doi: 10.1109/isbi.2019.8759531.

[15] N. S. K. M. K. Tirumanadham, T. S, and S. M, "Evaluating Boosting Algorithms for Academic Performance Prediction in E-Learning Environments," 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), pp. 1–8, Jan. 2024, doi: 10.1109/iitcee59897.2024.10467968.

[16] Q. Cheng et al., "A novel prognostic signature of transcription factors for the prediction in patients with GBM," Frontiers in Genetics, vol. 10, Oct. 2019, doi: 10.3389/fgene.2019.00906.

[17] J. Sheng, M. Shao, Q. Zhang, R. Zhou, L. Wang, and Y. Xin, "Alzheimer's disease, mild cognitive impairment, and normal aging distinguished by multi-modal parcellation and machine learning," Scientific Reports, vol. 10, no. 1, Mar. 2020, doi: 10.1038/s41598-020-62378-0.

[18] A. Sau and I. Bhakta, "Screening of anxiety and depression among seafarers using machine learning technology," Informatics in Medicine Unlocked, vol. 16, p. 100228, Jan. 2019, doi: 10.1016/j.imu.2019.100228.

[19] Koneru, S., Madhavi, P., Thota, K. K., Ramesh, J. V. N., Thatha, V. N., & Praveen, S. P. (2024). Employing Deep Learning Techniques for the Identification and Assessment of Skin Cancer. *Fusion: Practice & Applications*, *16*(2).

[20] Voddi, S., Sirisha, U., Praveen, S. P., Pandraju, T. K. S., Al-Dmour, N. A., & Islam, S. (2024, December). Hybrid CNN-GCN Model for Tumor Classification: Integrating Spatial Relationships in Medical Imaging. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-6). IEEE.

[21] Ponnaganti, N. D., Kumar, T. K. M., Praveen, S. P., Sindhura, S., Al-Dmour, N. A., & Islam, S. (2024, December). A Robust SVM Framework for Heart Disease Detection Utilizing Advanced Feature Selection Techniques. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-7). IEEE.

[22] PBV, R. R., Kumar, T. K. M., Praveen, S. P., Sindhura, S., Al-Dmour, N. A., & Islam, S. (2024, December). Optimizing Lung Cancer Detection: The Synergy of Support Vector Machine and Random Forest. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-7). IEEE.

[23] DONEPUDI, S., SIRISHA, G., & PAPPULA MADHAVI, S. P. (2024). OPTIMIZING DIABETES DIAGNOSIS: ADGB WITH HYPERBAND FOR ENHANCED PREDICTIVE ACCURACY. *Journal of Theoretical and Applied Information Technology*, *102*(23).

[24] PRAVEEN, S. P., ANUSHA, P. V., AKARAPU, R. B., KOCHARLA, S., PENUBAKA, K. K. R., SHARIFF, V., & DEWI, D. A. (2025). AI-POWERED DIAGNOSIS: REVOLUTIONIZING HEALTHCARE WITH NEURAL NETWORKS. *Journal of Theoretical and Applied Information Technology*, *103*(3).

[25] Praveen, S. P., Saripudi, V., Harshalokh, V., Sohitha, T., Karthik, S. V. S., & Sreekar, T. V. P. S. (2023, December). Diabetes Prediction with Ensemble Learning Techniques in Machine Learning. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1082-1089). IEEE.