SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V11I11P102

Volume 11 Issue 11, 11-19, November 2024
© 2024 Seventh Sense Research Group®

Original Article

Early-Stage Detection of Alzheimer's Disease Using Modified Firefly Based Ensemble Classification Model

Sri Lakshmi[1], Sreenu Babu[2]

[1,2]*MVR College of Engineering and Technology, Andhra Pradesh, India.*

[1]*Corresponding Author : srilakshmid2000@gmail.com*

Received: 26 September 2024 Revised: 22 October 2024 Accepted: 10 November 2024 Published: 22 November 2024

Abstract - *Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline affecting millions of people globally. Early detection is crucial to managing the disease effectively, enabling timely intervention to slow its progression. Most conventional machine learning approaches have been explored to enhance early prediction; however, these techniques often suffer from issues like overfitting, complex feature selection, and the inability to handle large datasets efficiently. To address the challenges of early AD prediction in the proposed model, the Hybrid Firefly optimized feature selection along with One Class SVM was used to achieve a better quality of Biomarker data and further integrate it with a customized ensemble classification model to improve the prediction rate. The proposed model performs better than conventional models and is measured in terms of accuracy, recall, precision and F1 score.*

Keywords - *Modified Firefly algorithm, One Class SVM, Ensemble Classification, XGboost, Feature Selection.*

1. Introduction

Alzheimer's disease is a progressive neurological disorder that leads to confusion, mood changes, and difficulty with daily activities. It is the most common form of dementia, accounting for 60-80% of dementia cases. Alzheimer's disease is a widespread neurodegenerative disorder affecting millions worldwide. It causes progressive damage to cognitive function, memory, and thinking abilities, devastatingly affecting individuals and society. MRI scans reveal the insidious nature of the disease, posing a significant threat to mental well-being. Early detection and intervention are crucial to mitigate its impact and improve the lives of those affected [1]. Genetic inheritance plays a role in up to 5% of Alzheimer's Disease (AD) cases. Specific gene mutations, such as APP, APOE, PSEN1, and PSEN2, have been linked to AD. These genetic associations were first identified between 1987 and 1993. Gene expression is a crucial process that converts genetic information into functional products. It transcribes genes into messenger RNA (mRNA) and then directs protein synthesis. These proteins perform various cellular functions, influencing disease susceptibility and progression [10]. Alzheimer's Disease (AD) is the most common neurodegenerative disorder, characterized by β-amyloid aggregation, tau hyperphosphorylation, synaptic dysfunction, and neuronal loss. Early detection biomarkers can enable timely treatment, improving disease management and patient outcomes. Biomarker development is crucial for effective AD diagnosis and treatment [9]. Mild Cognitive Impairment (MCI) is an intermediate stage in the Alzheimer's disease continuum, between age-related decline and dementia. Cognitive issues become noticeable and detectable through tests, but daily life activities remain unaffected. MCI is a critical transitional phase in the progression towards dementia. Research in machine learning and deep learning for Alzheimer's disease classification combines multiple modalities like MRI, PET, and CT scans. Clinical information like demographics, cognitive scores, and genetics is integrated with CNN-based classifiers.

This approach enhances classification accuracy and improves diagnosis. Advancements in AD classification hold promise for better treatment [13]. This degeneration leads to motor symptoms such as tremors, slow movement, and muscle stiffness. Non-motor symptoms like sleep disturbances, anxiety, and constipation also occur. PD's progressive nature significantly impacts the quality of life for those affected [8]. Following feature selection, two classification techniques were employed: Artificial Neural Networks (ANN) and Deep Neural Networks (DNN). This study proposes two hybrid diagnostic systems: GA_ANN and GA_DNN. These systems aim to predict dementia and its risk factors. The GA_ANN model combines genetic algorithms with (ANN), while the GA_DNN model integrates genetic algorithms with (DNN). These hybrid models leverage the strengths of both techniques to improve diagnostic accuracy [12].

Machine learning techniques often face challenges with high-dimensional data, making it difficult to identify the optimal feature set. Many features in a dataset are redundant or irrelevant, causing redundancy and irrelevancy issues. This leads to decreased model performance and increased complexity. Feature selection techniques help mitigate these issues by selecting the most informative features [6]. Silvia Basaia et al. (2019) developed a deep learning algorithm using Convolution Neural Networks (CNNs) to predict Alzheimer's disease (AD) diagnoses.

The algorithm achieved exceptional accuracy (above 98%) in distinguishing AD from healthy controls (HC) using the ADNI and combined ADNI + Milan datasets. It also showed promising results in differentiating mild cognitive impairment (c-MCI) from HC, with accuracy up to 86%. This research highlights the potential of CNNs in AD diagnosis, demonstrating high performance and clinical applicability. The algorithm's accuracy indicates its potential for accurate AD diagnosis [3]. Genotype datasets in Genome-Wide Association Studies (GWAS) can be vast, containing up to a million SNPs and a few thousand samples. Using such data directly for Machine Learning (ML) classification can lead to overfitting. Overfitted models perform well on training data but poorly on new data.

Feature selection and dimensionality reduction techniques can help overcome this challenge [11]. A recent study developed a comprehensive gene selection pipeline, integrating filter, wrapper, and unsupervised methods, to identify key features contributing to Alzheimer's Disease (AD). This approach combined multiple techniques to select relevant genes. Similarly, an ensemble method based on consensus-guided unsupervised feature selection was designed to detect genes associated with Huntington's disease. This approach leveraged the strengths of multiple methods to identify disease-related genes. The study demonstrated the effectiveness of combining different techniques for gene selection in neurodegenerative diseases [5].

This study focuses on comparing and identifying effective classification methods for a relatively small dataset. To achieve this, we implemented and compared three widely used classifiers: Support Vector Machine (SVM), Random Forest, and Extreme Learning Machine (RELM). These classifiers were evaluated for their performance in classifying the dataset. The goal is to determine the most effective classifier for this specific dataset. By identifying the best classifier, we can improve the accuracy of predictions and decision-making [4]. As Alzheimer's disease (AD) progresses, other brain areas are also impacted. In these cases, whole-brain approaches are preferred to focus on the affected regions. This approach enables a more accurate definition of brain hypertrophy in AD and Mild Cognitive Impairment (MCI) cases. Whole-brain analysis provides a comprehensive understanding of brain changes. It helps identify patterns and changes that may not be apparent when focusing on a single region. This approach can lead to a better understanding of disease progression and improved diagnosis [2]. The Whole-Brain Optimization (WOA) approach was used to segment brain sub-regions and identify disease-related alterations. The Hybrid Whole-Brain Optimization (HWGO) technique was integrated to enhance WOA's performance. HWGO improves exploration and exploitation phases, enabling optimal solutions for large-scale problems in real time. This hybrid approach enhances disease identification and brain analysis capabilities [14].

2. Related Work

Murali Krishna et al. [1] a novel approach for Alzheimer's disease prediction, focusing ona proposed model leveraging the ADNI2 dataset. The methodology integrates a random forest classifier with a Modified Artificial Bee Colony (MABC) algorithm for feature selection to enhance prediction accuracy and reduce computational complexity. Despite its advancements, limitations such as data heterogeneity and generalizability remain pertinent challenges. The model achieves a notable accuracy of 98%, signifying significant progress in disease prediction efficiency, crucial for healthcare decisions in resource-constrained settings.

Mehrdad Rostami et al. [2] explore the critical role of feature selection in enhancing machine learning models dealing with high-dimensional datasets, where many features are often irrelevant or redundant. By reducing dataset size, feature selection decreases computational complexity and improves prediction models' accuracy. The review categorizes and compares various feature selection methods, focusing specifically on wrapper and filter methods based on swarm intelligence (SI) algorithms. It evaluates the strengths and weaknesses of these SI-based methods, analyzing factors contributing to their effectiveness in optimizing feature subsets for improved data mining tasks.

Afrah Salman Dawood et al. [3] introduce a novel Deep Convolutional Neural Network (DCNN) architecture incorporating variations of CNN, modified VGG-16, VGG-19, ResNet50, and DenseNet121, augmented by a newly integrated classification layer. This framework is designed for detecting and categorizing Alzheimer’s disease using image-based methodologies. While achieving promising results with CNN-based models showing the highest accuracy of 96% and minimal loss of 9.92%, the study underscores the necessity of addressing inherent limitations such as dataset diversity and generalizability across broader pathological categories. The scope of this research highlights significant advancements in leveraging ML and DL techniques within medical diagnostics, advocating for further refinement to enhance model precision and clinical applicability.

M. Sudharsan et al. [4] propose an early diagnostic method for Alzheimer's disease using Mild Cognitive Impairment (MCI) and Structural Magnetic Resonance (sMR) imaging. They employ an Import Vector Machine (IVM), Regularized Extreme Learning Machine (RELM), andSupport Vector Machine (SVM) for classification, utilizing a greedy score-based strategy for feature selection and a kernel-based approach for data transformation. The study, based on Alzheimer's Disease Neuroimaging Initiative (ADNI) data, demonstrates that RELM significantly improves classification accuracy across AD, MCI, and Healthy Control (HC) groups. However, the review identifies limitations, such as the scarcity of trained samples and the complexity of feature descriptions, which pose challenges in early diagnosis. Nevertheless, the promising accuracy results suggest the potential for enhancing Alzheimer's disease detection, highlighting the scope for further research in refining these machine learning models for clinical application.

Petros Paplomatas et al. [5] proposed ensemble feature selection methodology for scRNA-seq data in Alzheimer’s disease integrates multiple strategies to identify dominant differentially expressed genes (DEGs), aiming to uncover transcriptome biomarkers. While promising, limitations include potential bias from selection methods and computational intensity due to high-dimensional data. Its scope is enhancing precision diagnosis and therapeutics through precise gene expression profiling. However, accuracy hinges on robust validation and integration with clinical data to ensure biological relevance and clinical applicability across diverse disease contexts. Qasem Al-Tashi et al. [6] focus on multi-objective feature selection in machine learning, examining studies from 2012 to 2019. Methodologically, it systematically analyzes techniques and algorithms employed to address this complex problem, emphasizing critical evaluations of their effectiveness. The scope encompasses identifying optimal features to enhance accuracy using various multi-objective approaches. However, limitations are noted, as no definitive solution has been universally established, highlighting ongoing challenges in achieving comprehensive feature selection. Overall, thereview provides insights into current methodologies and identifies promising research avenues for future exploration in this field.

Ashir Javeed et al. [7] proposed a methodology integrating adaptive synthetic sampling tomitigate bias and novel feature extraction techniques optimized with Support Vector Machine (SVM) using radial basis function (rbf). Despite advancements, limitations suchas variable accuracy and inherent biases in machine learning models persist. The scope extends to enhancing early detection, which is crucial given that symptom onset is potentially a decade before clinical manifestation. The proposed FEB-SVM model demonstrates significant improvement with 93.92% accuracy, surpassing existing state-of-the-art models in dementia prediction.

Noushath Shaffi et al. [8] present a hybrid ensemble method for the early detection of Parkinson's disease (PD) and Alzheimer's Disease (AD) that combines KNN and SVM. This approach aims to improve classification accuracy and robustness by utilising the parametric benefits of SVM and the non-parametric strengths of KNN. The approach admits drawbacks such as dependence on feature selection and parameter tweaking, which might impair generalizability across different datasets and clinical contexts. However, it shows promisein obtaining improved accuracy and specificity equivalent to Deep Learning (DL) techniques. The approach's scope is limited to the ADNI, OASIS, and NTAU PD datasets, indicating its generalizability to other well-known AD and PD databases. Its promise as a workable substitute for DL techniques is highlighted by the stated accuracy, especially in healthcare settings where obtaining large-scale training data may be difficult or impossible.

Ana Gabriela Sanchez-Reyna et al. [9] propose a methodology for classifying subjects into Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer’s Disease (AD) using data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. The study begins with feature selection from 2163 features using the GALGO genetic package, resulting in four significant features. Four classification techniques—Logistic Regression (LR), Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Machines (SVM)—are then applied with a 70% training and 30% test cross-validation setup. LR emerges with the highest area under the curve (AUC) of 0.842, demonstrating its efficacy in distinguishing between CN and MCI/AD subjects. However, the study is limited by its reliance on a single database (ADNI), potentially affecting generalizability to broader populations and datasets. The scope of the research focuses on feature selection and classification within the ADNI cohort, highlighting LR's performance in this context but also suggesting the need for validation across diverse datasets and clinical settings to enhance robustness and applicability.

Hala Alshamlan et al. [10] focus on evaluating various feature selection methods combined with SVM classification to identify biomarker genes in Alzheimer’s disease (AD). The proposed methodology involves comparing mRMR and F-score methods, showing they achieve high accuracy (around 84%) with 20-40 genes from a dataset of 696 samples and 200 genes. Limitations include potential overfitting due to small sample size and generalizability concerns. The scope highlights the potential for more accurate AD diagnosis and treatment with these methods. Overall, the study underscores that mRMR and F-score are effective in biomarker identification, surpassing GA, Chi-Square Test, and CFS methods in accuracy and applicability to AD research.

Nicholas Pudjihartono [11] highlights the challenges in genotype-based disease risk prediction using machine learning, emphasizing the "curse of dimensionality" due to many genetic features relative to sample size. Feature selection methods are crucial to enhance model generalizability by identifying relevant SNPs while filtering out noise and redundancy. Existing methodologies vary in their approaches and effectiveness, necessitating a comprehensive evaluation of advantages and limitations. Proposed methodologies typically involve combinations of filter, wrapper, and embedded techniques tailored to specific datasets and disease contexts. Achieving high prediction accuracy hinges on the ability to balance feature informativeness with computational feasibility and robust model performance across diverse populations and genetic profiles.

Ana Luiza Dallora et al. [12] present a novel approach for predicting dementia ten years in advance using multifactorial data encompassing 75 variables. It employs genetic algorithms for feature selection, artificial neural networks, and deep neural network models for classification. The proposed model achieves significant performance metrics with an accuracy of 93.36%, outperforming 11 other machine learning techniques previously applied in dementia prediction. Key predictors identified include age, past smoking habits, history of infarction, depression, hip fracture, single leg standing test (right leg), physical component summary score, and history of TIA/RIND. However, the study is limited by potential biases in data collection and the need for validation across diverse populations. Nonetheless, its scope lies in enhancing early detection efforts to potentially mitigate the impact of dementia through targeted interventions.

Noushath Shaffi et al. [13] identify that AI-based approaches in CAD for AD face challenges such as data scarcity and the interpretability of DL models despite their ability to learn complex data rapidly. To address these, an ensemble ML classifier for MRI data achieved 96.52% accuracy, outperforming individual classifiers and some DL models, especially in limited labeled data scenarios. Evaluations using ADNI and OASIS datasets compared ML classifiers and CNN-centric DL algorithms, providing insights into their respective strengths and weaknesses aiding algorithm selection based on data availability.

Chitradevi Dhakhinamoorthy et al. [14] integrate the Whale Optimization Algorithm (WOA) and Gray Wolf Optimization (GWO) into a hybrid approach (HWGO) for segmenting brain sub-regions crucial for diagnosing Alzheimer’s disease. This hybrid method aims to overcome WOA's potential for local optima by leveraging GWO's complementary optimization strategy, achieving a segmentation accuracy of 92%. Despite its success, limitations may include the sensitivity of optimization outcomes to parameter settings and the computational complexity involved. Future research could explore enhancements in scalability and robustness, potentially extending its application beyond brain imaging to broader medical image analysis tasks. The combined approach's high accuracy of 90% in AD classification showcases its effectiveness when paired with deep learning, underscoring its promising role in clinical diagnostics.

Mohamadreza Khosravi et al. [15] integrates a novel CAM-CNN model for enhancing the early detection of Alzheimer’s disease (AD) through MRI imaging. This approach combines a convolutional neural network with a cascade attention mechanism, leveraging two constraints, cost functions and cross-network diversity, to improve classification accuracy and processing efficiency. New cost functions, Satisfied Rank loss and Cross-network Similarity loss enhance network collaboration and performance robustness.

Validation on the Kaggle dataset demonstrates a high accuracy of 99.07% in multiclass AD classification, showcasing its potential for precise disease subtype detection and clinical utility. However, limitations may arise from dataset biases or generalizability to diverse populations. This warrants further exploration in real-world clinical settings to ascertain its broader applicability and accuracy in varied patient cohorts.

Ranjan Kumar et al. [16] utilize the Influenza Research Database and Human Surveillance Records for data analysis, employing ensemble-based stacked algorithms to enhance predictive accuracy. Evaluation metrics such as sensitivity and specificity validate the proposed efficient models, showcasing their potential as cost-effective tools for rapid influenza diagnosis across diverse populations.

Sina Fathi et al. [17] propose an ensemble deep learning method for early Alzheimer’s disease diagnosis using MRI images, detailing dataset collection, preprocessing, and model creation, including six top-performing CNN-based classifiers. Evaluation of ADNI data demonstrated high accuracies across six classification groups. Local dataset validation indicated 88.46% accuracy for three-way classification, outperforming individual models and aligning with top-tier studies. Limitations include the need for broader dataset validation. At the same time, the method's scope lies in its potential for enhancing Alzheimer’s diagnosis with deep learning techniques, showcasing competitive accuracy in early detection scenarios.

Aliaa El-Gawady et al. [18] propose a structured approach to predict Alzheimer’s disease (AD) from Gene Expression (GE) data, employing preprocessing, hybrid Gene Selection (GS), and classification with three ML models. It achieves high-performance metrics with MLP and SVM classifiers, which show notable effectiveness. Limitations include data scarcity in AD research, suggesting future work focuses on broader datasets for enhanced generalizability.

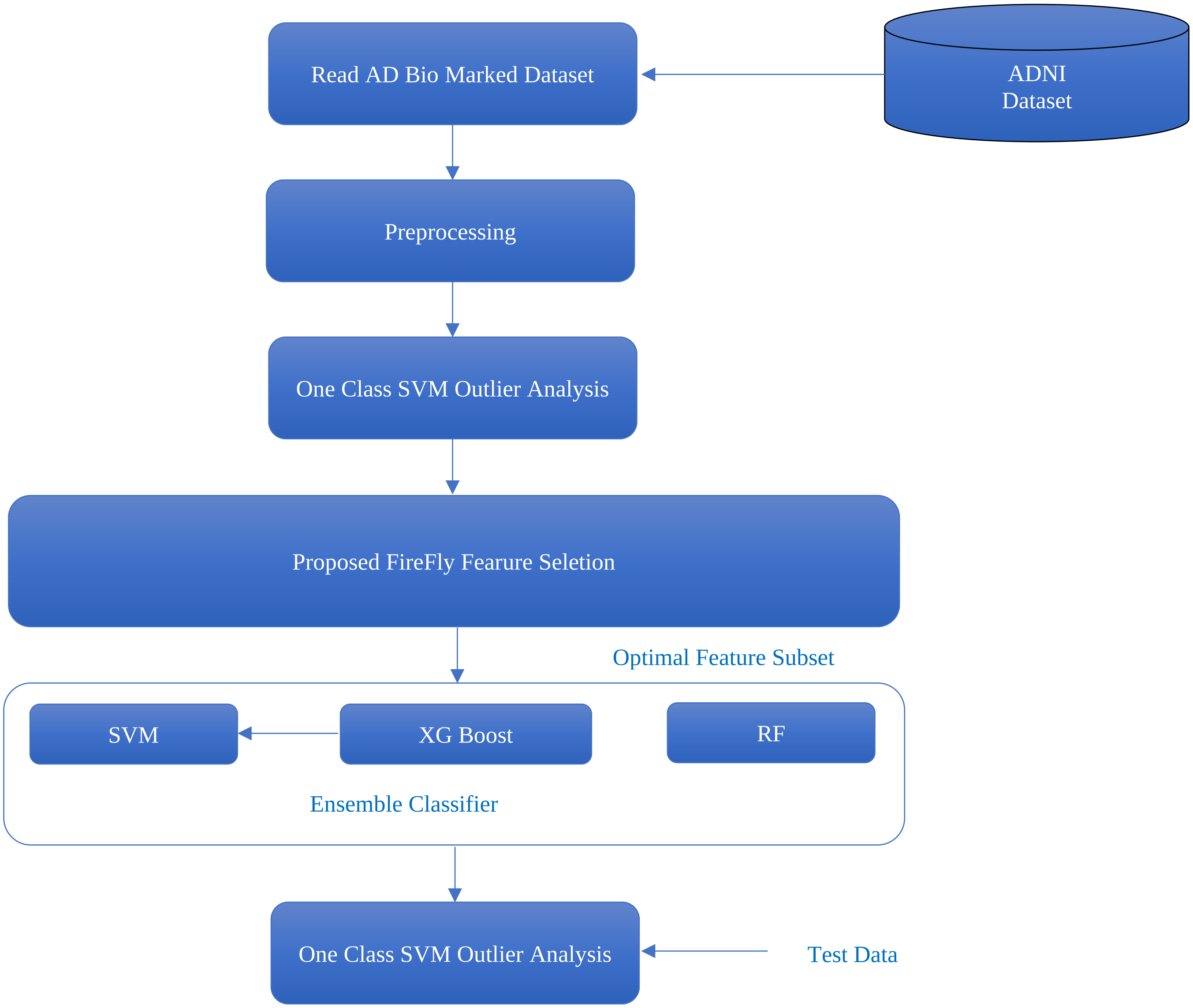

Read AD Bio Marked Dataset
ADNI
Dataset
Preprocessing
One Class SVM Outlier Analysis
Proposed FireFly Fearure Seletion
Optimal Feature Subset
SVM
XG Boost
RF
Ensemble Classifier
One Class SVM Outlier Analysis
Test Data

Fig. 1 Proposed model

3. Materials and Methods

The proposed model uses an ensemble classification model with modified Firefly Feature selection to reduce the problem of local optima in the evolutionary algorithm in feature selection over high dimensional datasets. Initially, the high-dimensional biomarker data set will be collected from ADNI (https://adni.loni.usc.edu/). Figure 1 refers to the concepts of the proposed model to improve the accuracy of early AD Prediction the model using the following stages.

1. Data Filtering
2. Anomaly Analysis
3. Proposed modified Firefly feature selection
4. Proposed Ensemble classifier

3.1. Stage 1 - Data Filtering

In the Bio marker-based ADNI data sample, most of the data is sparse, irrelevance between the features, and most continuous attributes have <INF> values. The preprocessing steps include label encoding, which converts categorical variables into numerical forms, and imputing missing values using the median of each column using data analysis.

3.2. Stage 2 - Anomaly Analysis using One Class SVM

A One-Class Support Vector Machine (SVM) is used to spot data points that significantly deviate from the norm, helping to uncover potential anomalies or errors that could impact further analysis.

One-Class SVM Algorithm
Step 1: Data Preparation
Collect Data: Gather a dataset consisting primarily of normal (non-anomalous) examples.
Normalize Data: Normalize or standardize the data to have zero mean and unit variance.

Step 2: Define Hyperparameters
Kernel Function: Choose a kernel function (e.g., linear, polynomial, RBF).
Kernel Parameters: Set parameters for the kernel function γ, Nu and ν parameter, which defines the upper bound on the fraction of margin errors and the lower bound of the fraction of support vectors.

Step 3: Train the Model by initialising the One-Class SVM model with the chosen kernel and parameters, then fit the trained model using the normal data.

Step 4: Predict
Predict on Training Data: Use the trained model to predict labels for the training data to evaluate model performance.
Predict on New Data: Use the trained model to predict labels for new data points. Classify them as normal or anomalous.

Step 5: Evaluate Model
Calculate Metrics: Calculate evaluation metrics (e.g., precision, recall) to assess the model's performance.
Tune Hyperparameters: If necessary, tune hyperparameters and re-train the model to improve performance.

3.3. Stage 3 - Proposed modified Firefly feature selection

The Firefly Algorithm is employed to select the most relevant features for Alzheimer's disease classification. This algorithm optimizes feature subsets to enhance classification accuracy by iteratively adjusting based on optimization criteria. This process is essential for managing high-dimensional data, as it reduces computational complexity and improves model interpretability by focusing on the most pertinent features.

Algorithm Firefly (n, MaxGen)

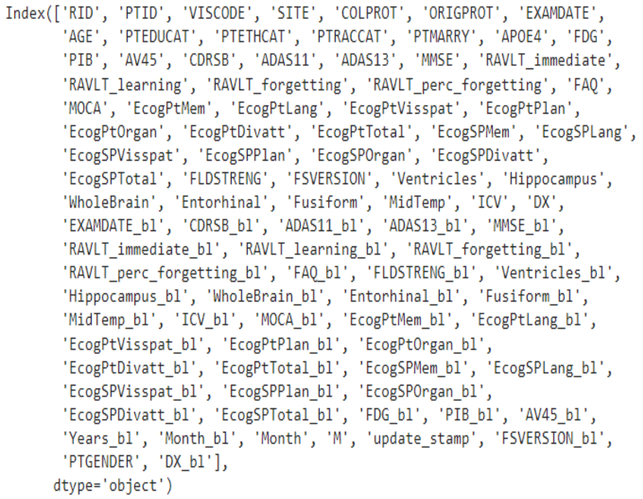
```
Define the objective function f(x), x = T ($x_1$... $x_a$);
Let us consider n fireflies
Generate an initial population of n fireflies $x_i$ (i = 1, 2, ..., n);
Light intensity $L_i$ at $x_i$ is determined by f($x_i$);
Define light absorption coefficient y;
while (t < MaxGen)
    for i: = 1 to n do
        for j = 1 to n do
            if ($L_i$ < $L_j$)
                for each firefly $x_i$ in x
                    for each firefly $x_k$ in $x_i$ with I($x_i$) < I($x_k$)
                        $x_i$ += $\beta$($x_k$ - $x_i$) + $\alpha$(rand - 0.5)
                        I($x_i$) = f($x_i$)
                    end_for
                end_for
               g* = find_global_best()
          end_for
          y = update_absorption_coefficient (y) (optional)
     end_for
end_while
Post-process results and visualization.
```

3.4. Stage 4 - Proposed Ensemble Classifier for Early AD Predication

The selected features are used to train a Random Forest classifier, chosen for its effectiveness in handling high-dimensional data and capturing complex feature relationships. Additionally, a Voting Classifier is applied, integrating predictions from SVM, Random Forest, and XGBoost classifiers to further boost classification performance. The Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) is a groundbreaking endeavour that converges the expertise of medical specialists, neurologists, and researchers to meticulously chronicle the progression of Alzheimer's disease. This ambitious project seeks to develop innovative early diagnostic tools and evaluation methodologies for Alzheimer's, leveraging a multifaceted approach incorporating genetic markers, advanced imaging processing techniques, and biological biomarkers. Through a longitudinal survey design, ADNI2 aims to create a comprehensive dataset comprising 94 distinct attributes, including genetic data and a range of clinical metrics, presenting a unique opportunity to investigate the correlations between specific genetic variants and cognitive decline in Alzheimer's patients. Ultimately, the ADNI2 research project endeavours not only to identify reliable indicators for early detection but also to establish a nuanced understanding of disease progression and its severity over time, thereby informing the development of more effective therapeutic strategies to combat this debilitating condition.

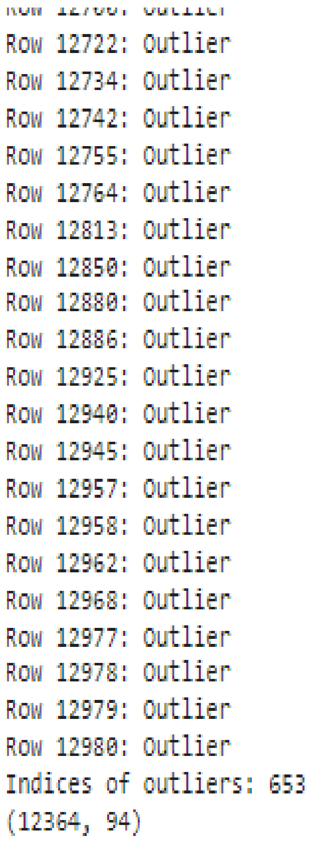
4. Results and Discussion

The proposed model, implemented and evaluated using Python libraries, exhibits enhanced performance characteristics relative to conventional classification approaches. A comprehensive assessment of model performance, encompassing accuracy, specificity, recall, precision, and F1 score, reveals notable advancements in predictive accuracy and computational efficiency. These findings suggest that our model offers a more effective and efficient solution for classification problems. A pivotal limitation of the traditional firefly algorithm lies in its propensity for local optima convergence. To address this shortcoming, our proposed model presents a modified firefly algorithm that leverages stochasticity to foster exploration and prevent premature convergence. By randomly selecting multiple partners and facilitating the exchange of optimal solutions, the algorithm iteratively refines its feature set, thereby minimizing the risk of local optima entrapment and enhancing the efficacy of the optimization process. The model was evaluated on test data, yielding more accurate results than traditional machine learning-based Anomaly Detection (AD) models. Notably, increasing the number of trees in the forest further enhanced performance. To optimize disease prediction accuracy, the proposed model employs an iterative approach, leveraging convergence to refine its predictions and achieve improved outcomes. Figure 2 refers to the overall features of the Biomarker AD dataset.



```
Index(['RID', 'PTID', 'VISCODE', 'SITE', 'COLPROT', 'ORIGPROT', 'EXAMDATE',
       'AGE', 'PTEDUCAT', 'PTETHCAT', 'PTRACCAT', 'PTMARRY', 'APOE4', 'FDG',
       'PIB', 'AV45', 'CDRSB', 'ADAS11', 'ADAS13', 'MMSE', 'RAVLT_immediate',
       'RAVLT_learning', 'RAVLT_forgetting', 'RAVLT_perc_forgetting', 'FAQ',
       'MOCA', 'EcogPtMem', 'EcogPtLang', 'EcogPtVisspat', 'EcogPtPlan',
       'EcogPtOrgan', 'EcogPtDivatt', 'EcogPtTotal', 'EcogSPMem', 'EcogSPLang',
       'EcogSPVisspat', 'EcogSPPlan', 'EcogSPOrgan', 'EcogSPDivatt',
       'EcogSPTotal', 'FLDSTRENG', 'FSVERSION', 'Ventricles', 'Hippocampus',
       'WholeBrain', 'Entorhinal', 'Fusiform', 'MidTemp', 'ICV', 'DX',
       'EXAMDATE_bl', 'CDRSB_bl', 'ADAS11_bl', 'ADAS13_bl', 'MMSE_bl',
       'RAVLT_immediate_bl', 'RAVLT_learning_bl', 'RAVLT_forgetting_bl',
       'RAVLT_perc_forgetting_bl', 'FAQ_bl', 'FLDSTRENG_bl', 'Ventricles_bl',
       'Hippocampus_bl', 'WholeBrain_bl', 'Entorhinal_bl', 'Fusiform_bl',
       'MidTemp_bl', 'ICV_bl', 'MOCA_bl', 'EcogPtMem_bl', 'EcogPtLang_bl',
       'EcogPtVisspat_bl', 'EcogPtPlan_bl', 'EcogPtOrgan_bl',
       'EcogPtDivatt_bl', 'EcogPtTotal_bl', 'EcogSPMem_bl', 'EcogSPLang_bl',
       'EcogSPVisspat_bl', 'EcogSPPlan_bl', 'EcogSPOrgan_bl',
       'EcogSPDivatt_bl', 'EcogSPTotal_bl', 'FDG_bl', 'PIB_bl', 'AV45_bl',
       'Years_bl', 'Month_bl', 'Month', 'M', 'update_stamp', 'FSVERSION_bl',
       'PTGENDER', 'DX_bl'],
      dtype='object')
```

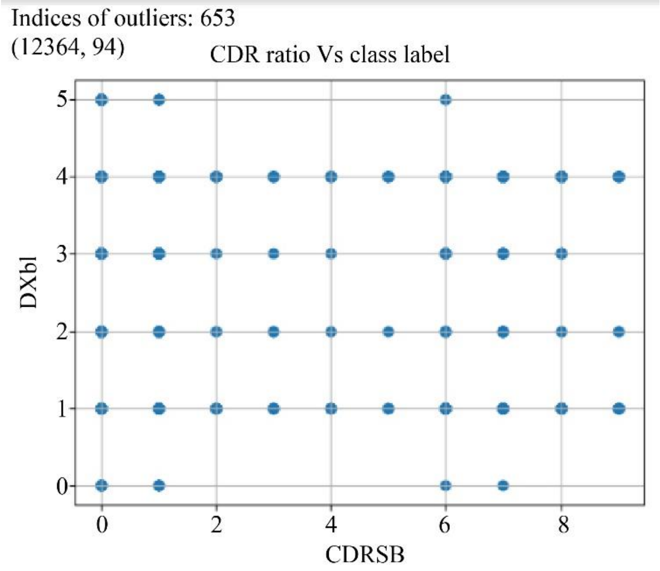
Fig. 2 Features within the High Dimensional Bio marker-based AD Data Set



```
Row 12722: Outlier
Row 12734: Outlier
Row 12742: Outlier
Row 12755: Outlier
Row 12764: Outlier
Row 12813: Outlier
Row 12850: Outlier
Row 12880: Outlier
Row 12886: Outlier
Row 12925: Outlier
Row 12940: Outlier
Row 12945: Outlier
Row 12957: Outlier
Row 12958: Outlier
Row 12962: Outlier
Row 12968: Outlier
Row 12977: Outlier
Row 12978: Outlier
Row 12979: Outlier
Row 12980: Outlier
Indices of outliers: 653
(12364, 94)
```

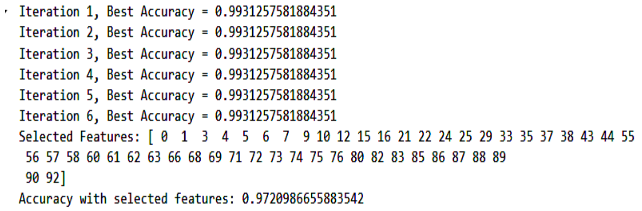
Fig. 3 Outlier Analysis and Correlation between the Features Data Set

Applied one class SVM model to raw data to achieve better quality, improve accuracy, and reduce the error rate in the early stage of AD prediction. The utilized data set in the proposed model totally consists of 13017 samples, and within these samples, 653 samples are identified as outliers. Figures 3 and 4 illustrate the overall concept of outlier analysis over a large set of features.



```
Indices of outliers: 653
(12364, 94)
```

Fig. 4 Outlier analysis features within the High Dimensional Bio marker-based AD Data Set

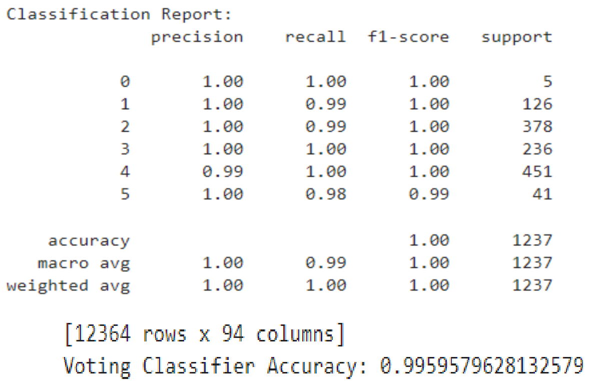


```
Iteration 1, Best Accuracy = 0.9931257581884351
Iteration 2, Best Accuracy = 0.9931257581884351
Iteration 3, Best Accuracy = 0.9931257581884351
Iteration 4, Best Accuracy = 0.9931257581884351
Iteration 5, Best Accuracy = 0.9931257581884351
Iteration 6, Best Accuracy = 0.9931257581884351
Selected Features: [ 0  1  3  4  5  6  7  9 10 12 15 16 21 22 24 25 29 33 35 37 38 43 44 55
 56 57 58 60 61 62 63 66 68 69 71 72 73 74 75 76 80 82 83 85 86 87 88 89
 90 92]
Accuracy with selected features: 0.9720986655883542
```

Fig. 5 Embedded-based firefly feature selection High Dimensional Bio marker-based AD Data Set

According to the iterative cycles report, the proposed model exhibits suboptimal performance, characterized by low accuracy, when the number of generated features falls below 42 out of the total 96 features. Conversely, the model achieves exceptional performance, with an accuracy of 97.209 %, in the 25th cycle, where a subset of 49 features is selected. This trend is consistently observed throughout the iterative process, suggesting that the optimal feature subset size for superior accuracy lies between 45 and 49 features. Figure 5 implies that the model's performance is significantly enhanced when a moderate to large number of features is selected rather than an excessively small or large subset.

The Classification Report presents the evaluation of our proposed classification model's performance using a classification report. Figure 6 notably shows that the model achieved its highest accuracy of 98.308% in the 25th cycle of model building, utilizing 49 biomarker features. This outstanding performance is further substantiated by the classification report's metrics, which reveal the number of correct estimations and those requiring correction. From Figure 7, the confusion matrix notably identified that 1232 test samples are correctly classified and 5 samples are misclassified, achieving the optimized accuracy of 99.59% compared to conventional models.



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      0.99      1.00       126
           2       1.00      0.99      1.00       378
           3       1.00      1.00      1.00       236
           4       0.99      1.00      1.00       451
           5       1.00      0.98      0.99        41

    accuracy                           1.00      1237
   macro avg       1.00      0.99      1.00      1237
weighted avg       1.00      1.00      1.00      1237

    [12364 rows x 94 columns]
    Voting Classifier Accuracy: 0.9959579628132579
```

Fig. 6 Embedded-based firefly feature selection High Dimensional Bio marker-based AD Data Set

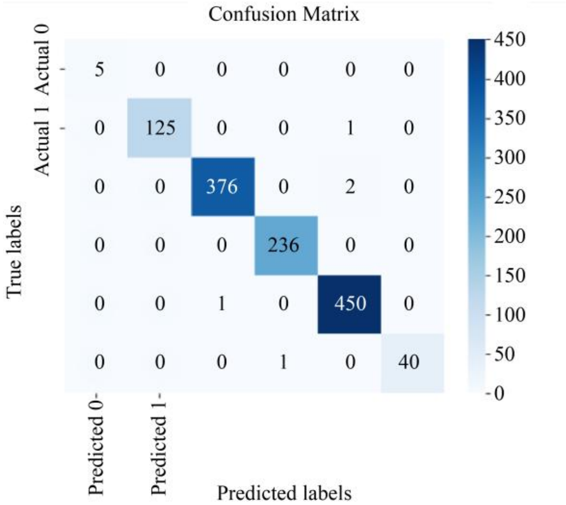


Fig. 7 Embedded-based firefly feature selection High Dimensional Bio marker-based AD Data Set

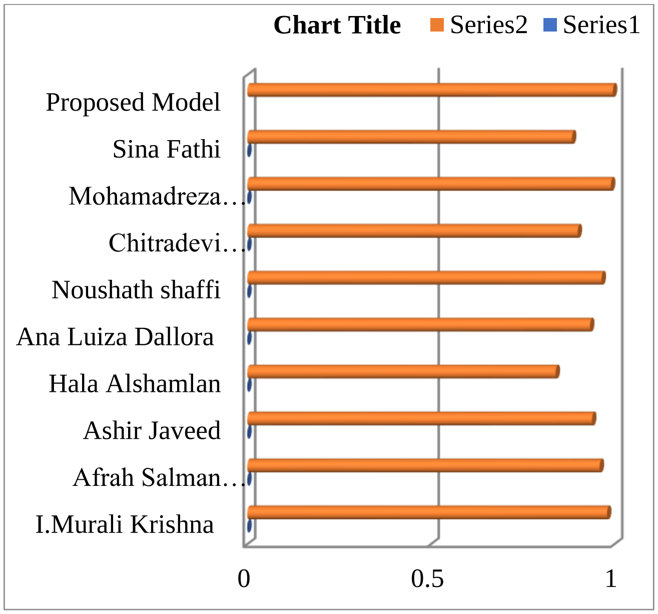


Fig. 8 Comparative study of the proposed model with conventional AD prediction

Table 1. Comparative analysis of the proposed model

Author	**Techniques Used**	**Accuracy**
I.Murali Krishna	Modified Artificial Bee Colony (MABC) Algorithm	98%
Afrah Salman Dawood	Deep CNN	96%
Ashir Javeed	Feature Extraction Battery (FEB), Support Vector Machine (SVM)	93.92%
Hala Alshamlan	SVM	84%
Ana Luiza Dallora	Genetic algorithm, Artificial Neural Network, Deep Neural Network	93.36%
Noushath shaffi	Ensemble Machine Learning classifiers, CNN-centric DL algorithms	96.52%
Chitradevi Dhakinamoorthy	Whale optimization algorithm, gray wolf optimization	90%
Mohamadreza khosravi	Novel CAM-CNN Model	99.07%
Sina Fathi	Ensemble Deep Learning method	88.46%
Proposed Model		99.59%

The comparative reports of the proposed model in Table 1 and Figure 8 and the conventional model enhance overall performance, scalability, and early diagnostic accuracy.

5. Conclusion

The proposed early-stage detection of Alzheimer's disease using the Modified Firefly-based ensemble classification model is capable of handling large and sensitive high-dimensional biomarker data. An innovative and nature-inspired optimization technique, Firefly Algorithm, can optimize feature selection in high-dimensional medical datasets, enhancing predictive accuracy while reducing computational complexity. By incorporating the Firefly Algorithm into the diagnostic process, this method aims to overcome the limitations of existing machine learning models, providing a more efficient, cost-effective, and reliable tool for early AD prediction.

The proposed model produced remarkable results, with an accuracy of 99.5%. The reason for its excellent performance could have been its capacity to record fine-grained features while maintaining spatial information.

Furthermore, to improve the model's efficiency, multiple random partners with GA are allowed, and it integrates with deep learning models to enhance overall performance and scalability. This approach is expected to improve early diagnostic accuracy and facilitate personalized treatment plans, helping to delay the progression of Alzheimer's and improve patients' quality of life.

References

[1] I. Murali Krishna et al., “Alzheimer's Disease Prediction Using an Enhanced Feature Selection of Multi Local Information Sharing Based on the ABC Algorithm,” *Journal of Data Acquisition and Processing*, vol. 39, no. 1, 2024. [Google Scholar]

[2] Mehrdad Rostami et al., “Review of Swarm Intelligence-based Feature Selection Methods,” *Engineering Applications of Artificial Intelligence*, vol. 100, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Afrah Salman Dawood, “A Comparative Study Using Deep Learning Models and Transfer Learning for Detection and Classification of Alzheimer’s Disease,” *Iraqi Journal of Computers, Communications, Control & Systems Engineering (IJCCCE)*, vol. 24, no. 1, pp. 57-70, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] M. Sudharsan, and G. Thailambal, “Alzheimer’s Disease Prediction Using Machine Learning Techniques and Principal Component Analysis (PCA),” *MaterialsToday Proceedings*, vol. 81, no. 2, pp. 182-190, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Petros Paplomatas et al., “An Ensemble Feature Selection Approach for Analysis and Modeling of Transcriptome Data in Alzheimer’s Disease,” *Applied Science*, vol. 13, no. 4, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Qasem Al-Tashi et al., “Approaches to Multi-Objective Feature Selection: A Systematic Literature Review,” *IEEE Access*, vol. 8, pp. 125076-125096, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ashir Javeed et al., “Early Prediction of Dementia Using Feature Extraction Battery (FEB) and Optimized Support Vector Machine (SVM) for Classification,” *Biomedicines*, vol. 11, no. 2, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Noushath Shaffi et al., “Bagging the Best: A Hybrid SVM-KNN Ensemble for Accurate and Early Detection of Alzheimer’s and Parkinson’s Diseases,” *Brain Informatics:16th International Conference*, Hoboken, NJ, USA, pp. 443-455, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ana Gabriela Sanchez-Reyna et al., “Feature Selection and Machine Learning Applied for Alzheimer’s Disease Classification,” *VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering*, Cancún, México, pp. 121-128, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hala Alshamlan et al., “Identifying Effective Feature Selection Methods for Alzheimer’s Disease Biomarker Gene Detection Using Machine Learning,” *Diagnostics*, vol. 13, no. 10, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Nicholas Pudjihartono et al., “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” *Frontiers in Bioinformatics*, vol. 2, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ashir Javeed et al., “Predicting Dementia Risk Factors Based on Feature Selection and Neural Networks,” *Computers, Materials & Continua*, vol. 75, no. 2, pp. 2491-2508, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Noushath Shaffi et al., “Performance Evaluation of Deep, Shallow and Ensemble Machine Learning Methods for the Automated Classification of Alzheimer’s Disease,” *International Journal of Neural Systems*, vol. 34, no. 7, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Chitradevi Dhakhinamoorthy et al., “Hybrid Whale and Gray Wolf Deep Learning Optimization Algorithm for Prediction of Alzheimer’s Disease,” *Mathematics*, vol. 11, no. 5, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohamadreza Khosravi, Hossein Parsaei, and Khosro Rezaee, “Novel Classification Scheme For Early Alzheimer’s Disease (AD) Severity Diagnosis Using Deep Features of the Hybrid Cascade Attention Architecture: Early Detection of AD on MRI Scans,” *Tsinghua Science and Technology*, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ranjan Kumar et al., “Ensemble Learning-Based Early Detection of Influenza Disease,” *Multimedia Tools and Applications*, vol. 83, pp. 5723-5743, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sina Fathi et al., “A Deep Learning-Based Ensemble Method for Early Diagnosis of Alzheimer’s Disease Using MRI Images,” *Neuroinformatics*, vol. 22, pp. 89-105, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Aliaa El-Gawady, BenBella S. Tawfik, and Mohamed A. Makhlouf, “Hybrid Feature Selection Method for Predicting Alzheimer’s Disease Using Gene Expression Data,” *Computers, Materials & ContinuaTechScience Press*, vol. 74, no. 3, pp. 5559-5572, 2023. [CrossRef] [Google Scholar] [Publisher Link]