RESEARCH PAPER

Advancing Neurodegenerative Disease Prediction: Innovations in Feature Engineering and Machine Learning

Akhilesh Deep Arya[a], Sourabh Singh Verma*[a], Prasun Chakrabarti[b] and Rimpy Bishnoi[c]

[a]Manipal University Jaipur, Jaipur 303007, Rajasthan India
[b]Sir Padampat Singhania University, Udaipur 313601, Rajasthan, India
[c]Techno NJR Institute of Technology, Udaipur 313003, Rajasthan India
**sourabhsingh.verma@jaipur.manipal.edu*

Received: 17th Dec, 2025; Revised: 10th Feb 2026; Accepted: 15th Feb, 2026; Available Online: 30th March, 2026

ABSTRACT
Early and precise diagnosis of neurodegenerative (Alzheimer's) disease is essential for efficient intervention and disease management. To improve disease prediction using the OASIS dataset, this research investigates a methodical framework that combines several machine learning classifiers with sophisticated feature selection techniques. We compare three most used feature selection techniques for this research: Correlation-Based Feature Selection (CFS), Wrapper Forward Selection (WFS), and LASSO regression with five classifiers: C5.0, CHAID, Logistic Regression, K-Nearest Neighbors (KNN), and Linear Support Vector Machine (LSVM). We also examine how the Synthetic Minority Oversampling Technique (SMOTE) overcomes the challenge of class imbalance. According to experimental results, LASSO and SMOTE combined with LSVM and CHAID produce better predictive performance, with accuracies of up to 94.66% and 94.34%, respectively. Interestingly, C5.0 achieves a 96.10% peak accuracy.
Index Terms: *Dementia, Machine Learning, Feature Importance, Neurodegenerative disease, Alzheimer's*
How to cite this article: Arya AD, Verma SS, Chakrabarti P and Bishnoi R, Advancing Neurodegenerative Disease Prediction: Innovations in Feature Engineering and Machine Learning. Int J Drug Deliv Technol. 2026;16(3): 290-303. DOI: 10.25258/ijddt.16.3.35
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

Neurodegenerative disease significantly reduces cognitive and functional abilities and in the later stage converted into Alzheimer's disease. Public health systems now prioritize early and precise diagnosis of dementia because the global burden will increase from 47 million affected individuals in 2020 to over 150 million by 2050 [1]. The lack of conclusive treatment after decades of research underlines the significance of early detection in halting the progression of the disease and enhancing patient outcomes.

Conventional diagnosis methods rest on clinical judgment and neuropsychological evaluations, which may not be sensitive and clearly visible in the early stages of dementia. By facilitating the automated analysis of intricate biomedical data, such as neuroimaging, genetic, and clinical parameters, modern developments in the area of machine learning (ML) present encouraging alternatives [2]. In high-dimensional datasets, ML techniques can discover subtle patterns and interactions that are frequently invisible using traditional statistical methods [3].

Using the OASIS dataset, a popular public clinical dataset, this study examines how feature engineering and model selection can increase the predictive accuracy of ML models for dementia diagnosis. We assess how the performance of five classifiers—Logistic Regression, CHAID, K-Nearest Neighbors (KNN), Linear Support Vector Machine (LSVM), and C5.0—is affected by three feature selection techniques: Wrapper Forward Selection (WFS), LASSO regression, and Correlation-Based Feature Selection (CFS) [4]. To handle the challenge of class imbalance, a prevalent issue in medical datasets, we also use the SMOTE.

This study aims to:

(1) To find the best feature selection and classification algorithm combination for the diagnosis of the neurodegenerative disease.

(2) Measure the impact of class rebalancing on model performance.

2. METHODS AND MATERIALS

The following section explains the experimental framework for the current researcher work. Figure 1 explains the step-by-step process, from dataset collection to the result analysis. The next section of the research article presents the results and analysis, where we discuss all the experiments performed using this setup and the outcomes obtained.

**Author for Correspondence: sourabhsingh.verma@jaipur.manipal.edu*

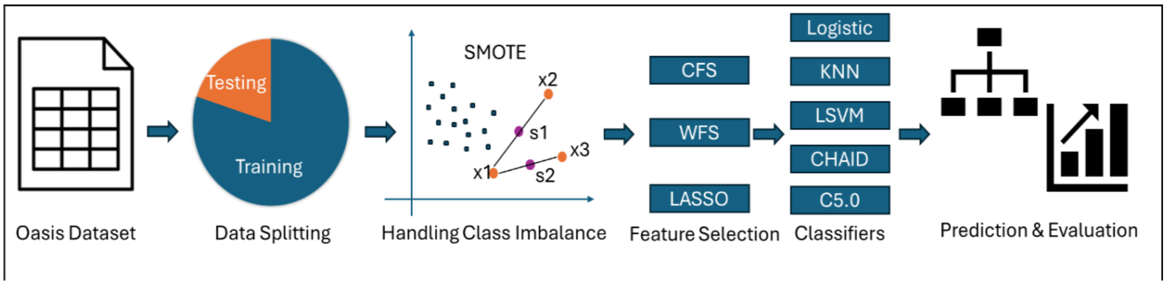

Figure-1 | Proposed experimental framework for neurodegenerative disease diagnosis.

2.1 Dataset:

The current research work uses the Open Access Series of Imaging Studies (OASIS) longitudinal dataset, which has 374 patient records with 15 characteristics explained in table-1 [5]. The target variable is a binary indicator that tells the difference between groups with dementia and groups without it.

Table-1 | Detail description of the OASIS dataset used for the research work

Sr. No.	Attribute Name	Description
1	SubjectID	Patients ID's
2	MRIID	MRI ID's
3	Group	Demented/ Normal
4	Indicator	0/1 indicating demented/ Normal Group
5	Visit	Visit Sequence (1-5)
6	MF	Male/ Female
7	Hand	Right/ Left
8	Age	Age of patient
9	EDUC	Education (6-23)
10	SES	Economic and Social Standing
11	MMSE	Mini Mental State Exam Scores
12	CDR	Clinical Dementia Rating
13	eTIV	Estimated total intracranial volume, mm3
14	nWBV	Normalized whole-brain volume
15	ASF	Atlas scaling factor

Before modeling, the dataset was cleaned to remove missing values, encoded correctly, and normalized so that all numerical features had the same scale.

2.2. Methodology Overview:

The proposed framework includes a multi-stage pipeline that combines feature selection, oversampling, classification, and evaluation, as shown below:

A. Dataset Preprocessing: Cleaning up the dataset to maintain the consistency and quality of the data.

B. Feature Selection: Identifying the most relevant features using CFS [6], Wrapper, and LASSO methods.

C. Handling class imbalance: Using SMOTE to fix class imbalance of the dataset, by generating some synthetic samples of the minority class (demented cases).

D. Training and Testing: We trained five machine learning classifiers—Logistic Regression, CHAID, K-Nearest Neighbors (KNN), Linear Support Vector Machine (LSVM), and C5.0—on both the original and balanced datasets. We use a number of metrics, such as accuracy, precision, recall, F1-score, AUC, and GINI coefficient to rate each model.

IBM SPSS was used for all of the experiments to create models and test their performance. Cross-validation was used to lower the chance of overfitting and make sure the results were strong.

2.3. Feature selection

Feature selection process identifies and uses only those input features that help in improving accuracy, reduce overfitting, and help model training in less time. It also increases the generalization of the data. This study uses three feature selection approaches, as stated below:

A. Correlation-based feature selection (CFS): picks feature that are vastly correlated with the target variable and have little redundancy with other features [6]. This makes the model easier to understand and works better.

B. Wrapper Method: Wrapper Forward Selection (WFS) adds features one at a time based on how much they improve the accuracy of the model [7]. This creates the best feature subsets for each algorithm, but it costs more to run.

C. Least Absolute Shrinkage and Selection Operator (LASSO): LASSO Regression uses L1 regularization to shrink the coefficients of less important features to zero. This selects features and reduces overfitting at the same time [8].

2.4. Classification Algorithms

We evaluate the following classifiers, the reason behind selecting these classifiers is they work well with medical tabular data in prediction of the disease:

A. Logistic Regression: It is generally used for binary classification. It is easy to understand and quick to compute. Logistic Regression can also be used for multi-class classification [9]. The simplicity and interpretability of logistic regression make it the favored method for medical diagnosis and disease prediction based on clinical data.

Logistic Regression models the probability that a given instance belongs to the positive class as:

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \dots \dots 1$$

- *Y:* The binary outcome variable, where *Y=1* indicates the dementia, and *Y=0* indicates a non demented case.
- $X = [X_1, X_2, \dots, X_n]$: The set of clinical features (e.g., age, cognitive test scores, SES, CDR and other biomarkers) for a given patient.
- β_0 The intercept term.
- $\beta_1, \beta_2, \dots, \beta_n$. The coefficients corresponding to each feature $X_1, X_2, \dots, X_n$ which are estimated from the training data [10]. A positive coefficient ($\beta_i > 0$) means that as a corresponding feature X_i increases the likelihood of the patient suffering from dementia (Alzheimer's) increases (e.g. Age). A negative coefficient means that as the corresponding feature increases the likelihood of the patient having dementia decreases (e.g. MMSE scores).
- The expression inside the exponent is:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n) \dots \dots \dots 2$$

It represents the log-odds of the probability of having dementia. Each coefficient βi represents the change in the log-odds of having Alzheimer's Disease for a one-unit change in the corresponding feature Xi, holding all other features constant.

B. Chi-squared automatic interaction detector (CHAID): CHAID (Chi-squared Automatic Interaction Detector) is a decision tree method that can find interactions between features. It works best with categorical predictors. It repeatedly divides the dataset according to chi-squared tests to determine interactions between features to maximize the split [11]. If no more splits are significant, CHAID ends here and produces a tree that is easy to interpret. Well-suited for healthcare due to its ability to process multi-level categorical predictors and an internal mechanism to reveal how features (e.g. clinical data) interact and contribute to the target variable.

For each categorial predictor variable, the chi-squared test statistics are computed as follows:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots 3$$

- O_{ij} = observed frequency in cell ij (e.g., the count of patients with a specific feature value and disease status) [12].
- E_{ij} = expected frequency in cell ij (based on the marginal totals).

C. Linear Support Vector Machine (LSVM): LSVM classifier works best with data that can be split into two groups, positive and negative [13, 14]. In our case it works by finding a hyperplane that maximizes the margin between demented and non-demented cases, providing us with a very robust decision boundary with convex optimization [15].

For a given OASIS dataset used, the objective function can be formulated as:

$$\min_{w,b} \frac{1}{2} ||w||^2 \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots 4$$

Subject to:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) \geq 1, \forall i \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots 5$$

- Where $\boldsymbol{w}$ is the weight vector that defines the orientation of the hyperplane.
- b is the bias term that shifts the hyperplane.
- $\boldsymbol{x_i}$ represents the feature vector of the $i - th$ sample (i.e. attributes like age, cognitive scores etc.).
- y_i is the class label of the $i - th$ sample where $y_i \in \{-1,1\}$. $y_i = 1$ means the sample is normal and $y_i = -1$ means the sample is demented

The hyperplane that separates the classes is given by:

$$w.x + b = 0 \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots 6$$

The margin between the classes is defined as:

$$Margin = \frac{2}{||\boldsymbol{w}||} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots 7$$

Thus, maximizing the margin is equivalent to minimizing $\frac{1}{2}||w||^2$, which is an objective function.

D. K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a non-parametric, distance-based algorithm that sorts instances based on how close they are to training samples [16].

An important point is the choice of k: a too-small k can be noisy and sensitive to outliers, a too-large k produces smooth decision boundaries. KNN is a very straight forward method that can work well for small to medium datasets as it uses distance metrics like Euclidean distance to find a point's neighboring points. While it is computationally exhaustive with large datasets, the simplicity and precision of KNN in multi-class classification and regression problems has made it a common tool for instance-based learning in many applications including pattern recognition and medical diagnosis [17].

Euclidean distance:

$$d(x_q, x_i) = \sqrt{\sum_{j=1}^{n} (x_{qj} - x_{ij})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

Where:

- x_q is the feature vector of the query point, representing the clinical information.
- x_i is the feature vector of the data point in the dataset.
- n is the number of features (clinical variables such as Age, MMSE, SSE etc.).
- j represents individual feature, and the summation computes the squared difference between each corresponding feature in x_q and x_i.

Identify the k closest neighbors to the query point xq based on the calculated distance matric [18]. These neighbors represent the k data points in the dataset that are most similar to xq.

Once the k closest neighbors are identified, the algorithm assigns a class label (demented, non-demented) to the query point xq based on the majority vote among the class labels of the k neighbors. If yi represents the class label of the $i - th$ neighbor [19]. The predicate class $\hat{y}q$ for the query point xq is:

$$\hat{y}_q = mode\ (y_1, y_2, \ldots\ldots\ldots, y_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9$$

Where mode refers to the most frequent class label among the k neighbors.

E. C5.0: C5.0 is a tree-based decision algorithm that supports boosting and reduces overfitting. It is known for being easy to understand and very accurate when working with structured data [20]. In contrast to linear models, which can be fitted using a straightforward mathematical formula, it is predicated on the construction of a decision tree that divides the data according to input feature values before recursively dividing those subsets [21].

Entropy and information gain:

- Entropy is a metric used to quantify data impurity or uncertainty. Entropy for a dataset S with classes C is calculated as:

$$H\ (S) = -\sum_{i=1}^{c} p_i\ log_2\ (p_i) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 10$$

Where pi is the proportion of instances belonging to class i in the dataset

- Information gain is used to decide the best features to split the data:

$$IG\ (S, A) = H\ (S) - \sum_{v\ \in Values\ (A)} \frac{|Sv\ |}{|S|} H\ (Sv) \ldots\ldots\ldots\ldots 11$$

Where A is an attribute, Values (A) are the possible values of attribute A, and Sv is the subset of S where attribute A has value v.

2.5. SMOTE

SMOTE is a technique employed to handle class imbalance in datasets. In practice, especially for medical diagnostics, there tend to be datasets with an imbalance between classes, with one class (e.g., patients with disease) is very less in comparison to another class (e.g., healthy people). Most machine learning algorithms are biased towards the majority class; hence, this imbalance would cause the model to underperform [22].

SMOTE solves this problem by producing synthetic instances for the lesser class instead of duplicating them. More specifically, for each data point selected from the minority class, the algorithm identifies its nearest neighbors and creates synthetic data points that lie along the line segments joining the selected data point with its neighbors [23]. Activated Class Path Rebalancing (ACPR) effectively up-samples the minority class in a manner that leads to a balanced dataset.

SMOTE improves the performance of classifiers significantly, especially in those that detect instances of the minority class. SMOTE can be useful in healthcare, where correctly identifying rare conditions (often represented by the minority class) is crucial, as it gives reservations to models that train data effectively by reducing bias. Data includes clinical variables, in our case it is age, MMSE scores, and diagnostic tests [24].

2.6. Performance Matrices

A. ***Accuracy***: Accuracy refers to the ability of a model to classify demented as well as non-demented cases in the context of AD diagnosis. It is computed as the fraction of correct classifications (including true positives and true negatives) to the total number of classifications produced by the model [25].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 12$$

B. ***Sensitivity***: Also known as True Positive Rate (TPR), Recall or Hit Rate, sensitivity measures how well a classification model classifies the demented cases as positive [25].

$$Sensitivity = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 13$$

C. Specificity: Specificity (or True Negative Rate (TNR)) is the measure of the True Negative Class, the proportion of normal or non-demented cases that are correctly classified as negative by a classification model, thus differentiating them from demented cases [26].

$$Specificity = \frac{TN}{FP + TN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 14$$

D. Precision: The ratio of true positive demented cases identified to the total number of cases predicted as positive (true and false positives) by the classification algorithm [26].

$$Precision = \frac{TP}{TP + FP} * 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots 15$$

E. F-Measure: In terms of a classification model's capacity to identify dementia, the F-measure (F1 score), which is the harmonic mean of precision and recall, offers a value that is balanced between the two metrics. This metric is especially useful for imbalanced classes and varying cost is associated with false positives and false negatives [27].

$$F - Measure = 2 * \frac{Precision * Recall}{Precison + Recall} \ldots\ldots\ldots\ldots 16$$

F. GINI coefficient: is a measure of how well a classification model separates normal cases from demented cases [28]. In the machine learning paradigm, it contributes to evaluating the model-fit to the data available (i.e., how well the model splits or makes decisions to discriminate between demented cases vs. non-demented cases).

G. Area under Curve (AUC): The AUC curve is a widely used graphical tool that shows how well a binary classification model can distinguish between a target class being present (demented) and the target class being absent (normal), as a function of the classification threshold. This is useful because it allows visualizing the model's ability to separate demented from non-demented across a spectrum of decision boundaries [29].

3. RESULT AND ANALYSIS

This section compares five machine learning classifiers—Logistic Regression, CHAID, LSVM, KNN, and C5.0—using different feature selection methods and SMOTE to balance the classes. We used stratified sampling to split the OASIS dataset into training and testing sets, which kept the class distributions the same. We then used several metrics, such as Accuracy, AUC, Sensitivity, Specificity, Precision, GINI coefficient, and F1-score, to measure how well the model worked.

3.1 Baseline Performance without Feature Selection

Table 2 shows how well the model works without selecting features. C5.0 had the best accuracy (95.23%), AUC (0.97), and GINI coefficient (0.96) of all the models, but it was less sensitive and precise than the others. CHAID and LSVM also did well, with the accuracy of 93.23% and 92.68%, respectively. Logistic Regression had average results, but KNN had the highest sensitivity (91.65%), even though its overall accuracy was lower.

Figure 2 is the comparison of Accuracy, sensitivity, specificity and precision of all classifiers. The area under the curve (AUC) comparison is shown in Figure 3, which shows C5.0 gets the best AUC of 0.97, and Logistic gets the lowest AUC of 0.88

Table-2 | Performance of different classifiers without feature selection

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
Logistic	89.87	90.00	90.80	91.94	0.88	0.79	91.08
CHAID	93.23	91.00	89.42	91.24	0.92	0.91	92.80
LSVM	92.68	90.83	90.44	92.41	0.89	0.78	91.61
KNN	81.75	91.65	91.94	92.44	0.94	0.92	92.61
C5.0	95.23	92.60	92.80	92.40	0.97	0.96	94.80

Insite: C5.0 can model complex interactions well even without explicit feature selection, probably because it can boost and use tree ensembles.

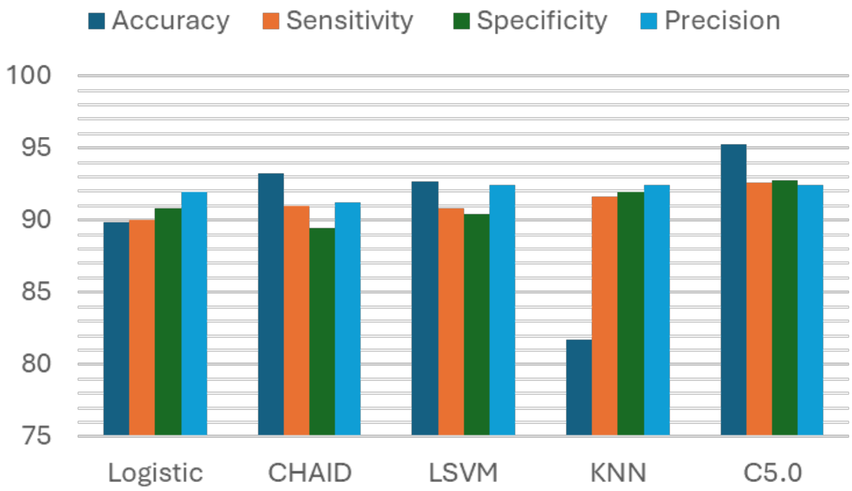

Figure-2 | All five classifiers' performance in the absence of feature selection

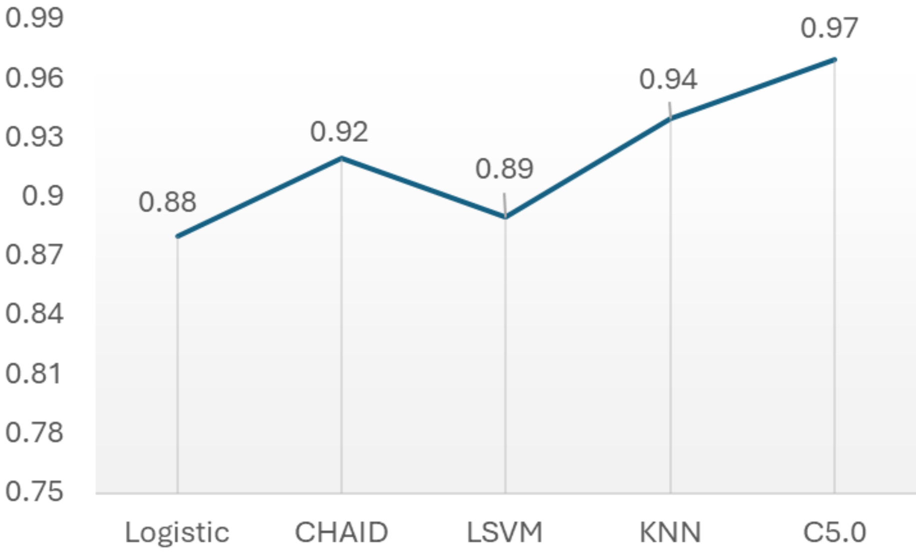

Figure-3 | AUC value comparison of all five classifiers without feature selection

3.2. Correlation-Based Feature Selection (CFS)

Using CFS made the model work better overall by getting rid of unnecessary features. Table-2 shows that LSVM had the best accuracy at 93.18%, as well as strong precision and F1-score. It was closely followed by C5.0 (93.36%) and KNN (92.78%). KNN had the best AUC (0.95) under CFS, which is interesting because it shows that it can separate things well even though it is simple.

Figure 4 is a comparison chart of all five classifiers with CFS as feature selection. Figure 5 shows AUC, KNN has the best AUC of 0.95. Figure 6 indicates the top three features which are contributing highest for prediction as per CFS and that are MMSE, SES, Age of patient for dementia prediction.

Table-3 | Performance of different classifiers with correlation- based feature selection

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
Logistic	91.87	91.65	91.40	88.94	0.90	0.92	92.80
CHAID	92.93	89.32	92.02	89.94	0.93	0.94	92.90
LSVM	93.18	91.44	93.35	93.44	0.92	0.90	93.61
KNN	92.78	91.34	93.24	93.94	0.95	0.94	94.51
C5.0	93.36	82.70	86.54	85.71	0.94	0.89	90.75

Key Finding: CFS was especially good at making LSVM and KNN better at generalizing, which supports the idea that decorrelated features make classifiers that work on margins or distances better.

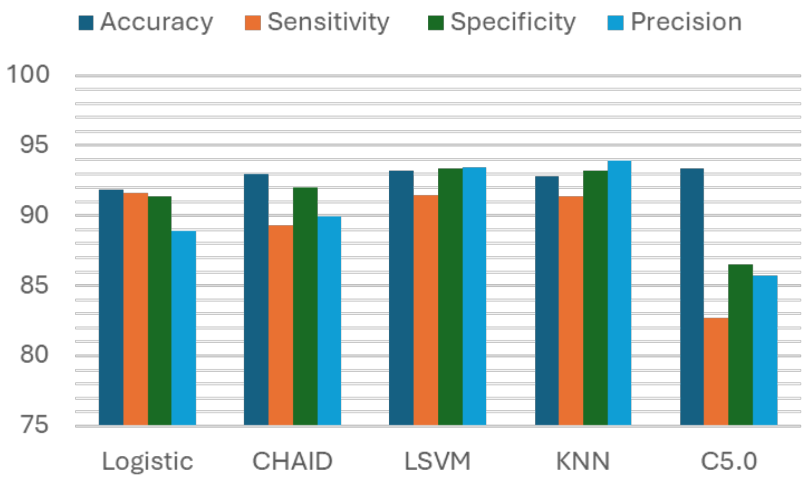

Figure-4 | Performance of all five classifiers with CFS

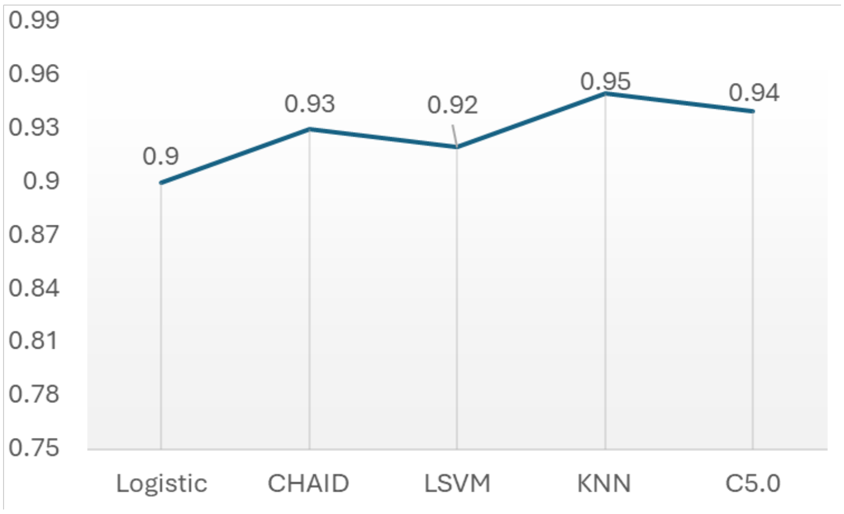

Figure-5 | AUC value comparison of all five classifiers with CFS

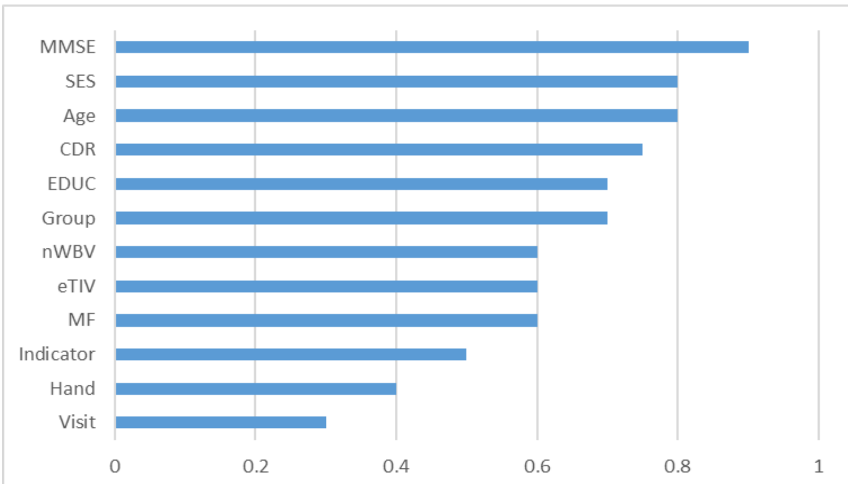

Figure-6 | Feature ranking by CFS

3.3. Wrapper Forward Feature Selection (WFS)

Table 4 reveals notable variances for each of the classification algorithms when using WFS. It gave the model performance the biggest boost. C5.0 had the highest accuracy ever recorded at 96.10%, as well as very high sensitivity (97.82%) and precision (98.55%). LSVM and KNN also did better, with both getting more than 92% accuracy.

A comparison chart of all five classifiers with the WFS method is depicted in Figure 7. The AUC was depicted in Figure 8, with C5.0 which gets a maximum value of AUC 0.95. As shown in Figure 9, the feature ranking by CFS shows that MMSE, Age and EDUC are the three most important features that contribute the most to prediction.

Table-4 | Performance of different classifiers with WFS

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
Logistic	89.00	91.34	91.80	90.87	0.89	0.88	99.27
CHAID	91.93	88.47	92.42	91.23	0.94	0.93	92.61
LSVM	93.68	90.36	92.44	91.87	0.92	0.87	92.30
KNN	92.75	93.22	90.94	92.78	0.91	0.93	95.85
C5.0	96.10	97.82	95.35	98.55	0.95	0.92	94.44

Observation: WFS fine-tunes feature subsets that are specific to model architecture. This makes it very useful for adaptive learners like C5.0 and non-parametric models like KNN.

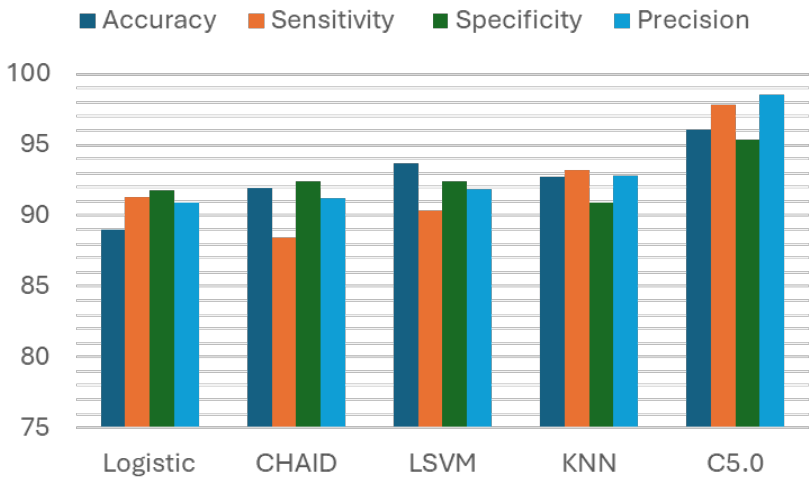

Figure-7 | Performance of all five classifiers with WFS

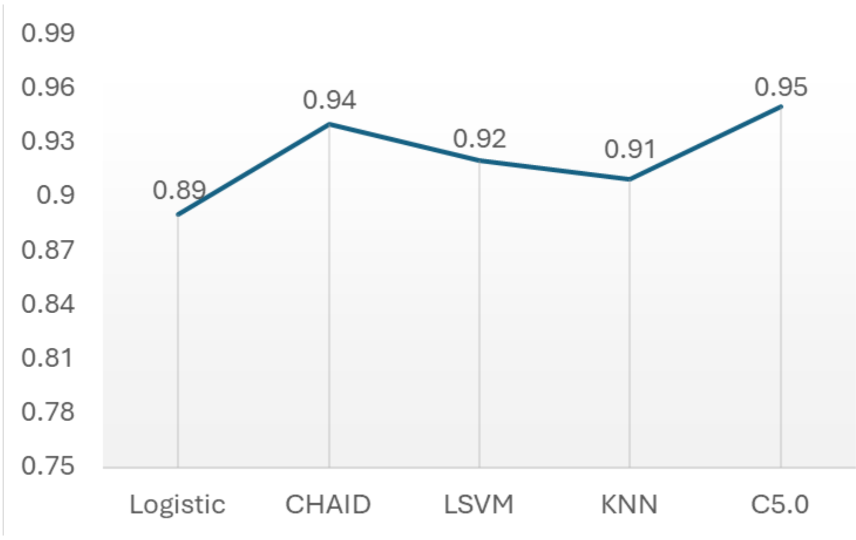

Figure-8 | AUC Value comparison of all five classifiers with WFS

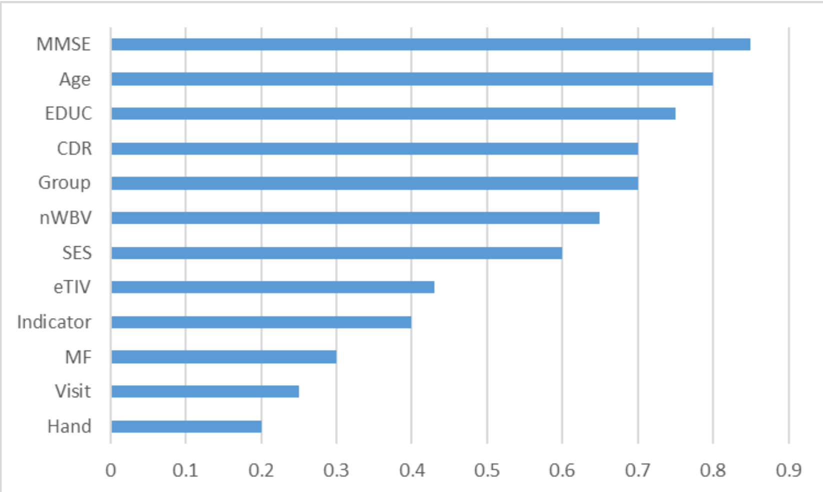

Figure-9 | Feature ranking by WFS

3.4. LASSO Feature Selection

Table 5 displays the performance matrix for each classifier using LASSO feature selection. LASSO gave very competitive results and also helped with model regularization. CHAID got 94.66% accuracy, the highest AUC (0.96), and a good balance between sensitivity and specificity. LSVM also did well all the time, with an accuracy of 94.34% and an F1-score of 95.89%.

Figure 10 shows a comparison of all five classifiers with the LASSO feature selection method. The AUC values are shown in Figure 11, where CHAID had the best AUC, with a value of 0.96, and then LSVM with AUC of 0.95. CFS is an embedded type of feature selection method, and the top three features with the highest prediction relevance are EDUC, MMSE, and Age as shown in Figure 12.

Table-5 | Performance of different classifiers with LASSO feature selection

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
Logistic	88.54	82.53	90.00	88.23	0.79	0.74	89.30
CHAID	94.66	95.21	98.74	98.56	0.96	0.94	95.83
LSVM	94.34	97.58	93.21	96.25	0.95	0.97	95.89
KNN	89.92	95.65	91.58	93.02	0.75	0.72	95.83
C5.0	95.63	88.23	91.25	89.93	0.91	0.82	96.00

Findings: LASSO's selection based on sparsity improved decision tree generalization (CHAID) and kept linear model performance (LSVM) with fewer features.

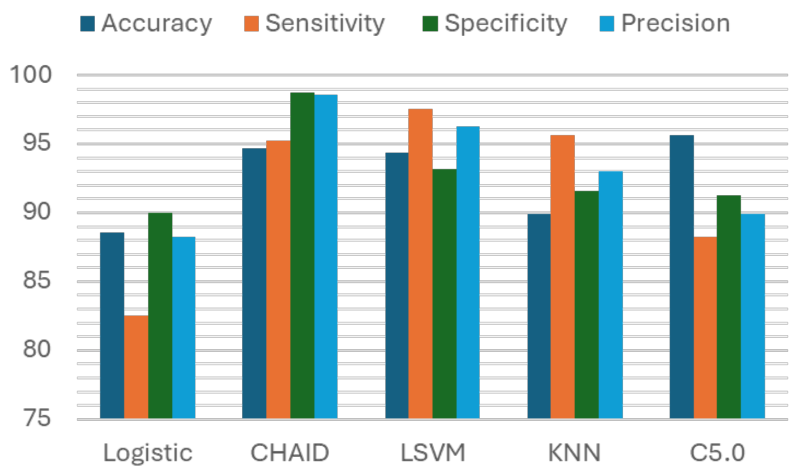

Figure-10 | All five classifiers' performance using LASSO feature selection

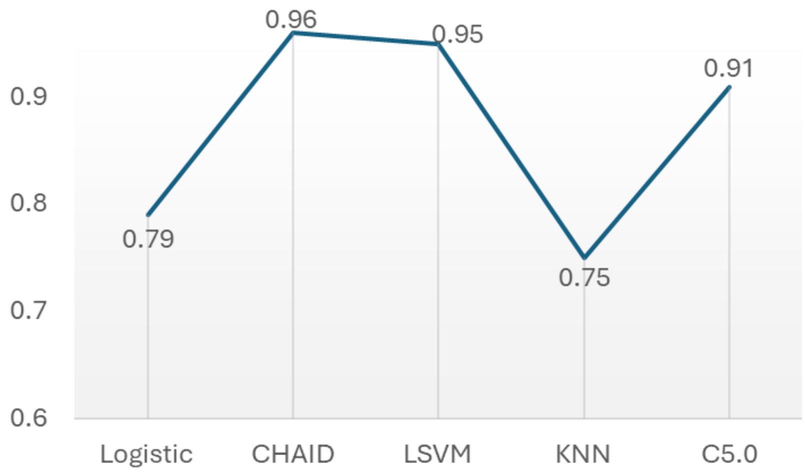

Figure-11 | AUC value comparison of all five classifiers with LASSO feature selection

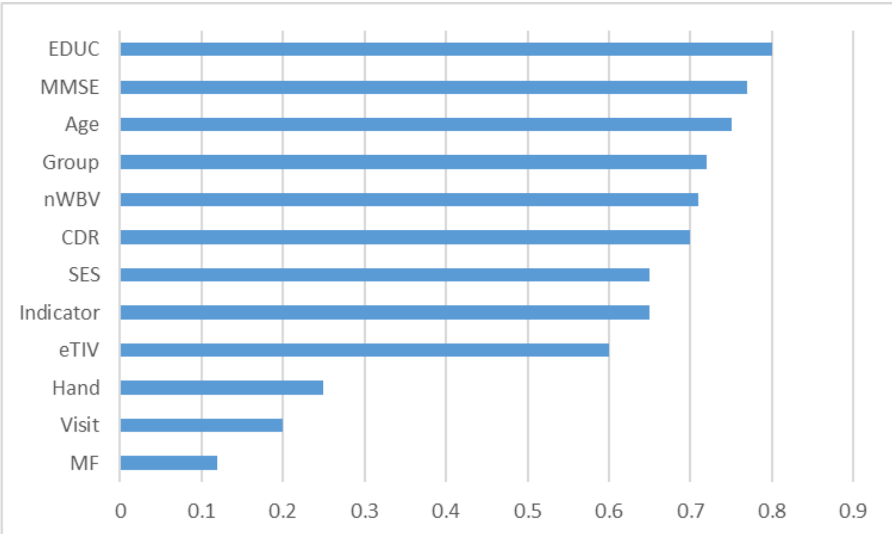

Figure-12 | Feature ranking with LASSO feature selection

3.5. Feature Selection with SMOTE

Logistic Regression consistently underperformed all feature selection methods. Therefore, we did not pursue logistic regression beyond initial exploration. Among feature selection methods, LASSO regression was more accurate than others. Thus, we used the features selected by LASSO for the other four classifiers. Table 6 shows performance matrix for the top 4 classifiers.

The best classification results came from using LASSO and SMOTE together. The best performers were LSVM and CHAID, which had 92.42% and 93.23% accuracy, respectively. CHAID also had the best AUC (0.99) and GINI (0.95) scores in all of the tests. C5.0 had the best accuracy (96.45%), but its F1-score and sensitivity were lower, which means it was overfitting to the majority class.

Figure 13 shows the results discussed above, using LASSO with top four classifiers and SMOTE. From the comparison in AUC scale, CHAID is higher than all four classifiers as depicted by Figure 14.

Table-6 | Performance of different classifiers with LASSO selected features and SMOTE

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
CHAID	93.23	92.36	96.00	95.94	0.99	0.95	97.62
LSVM	92.42	94.32	96.68	98.50	0.97	0.98	97.34
KNN	94.30	91.57	93.40	92.31	0.95	0.83	95.26
C5.0	96.45	90.00	88.23	82.32	0.98	0.94	86.00

Submission: SMOTE worked well to improve model recall for the minority (demented) class, and when used with LASSO, it made classifiers like LSVM and CHAID work better.

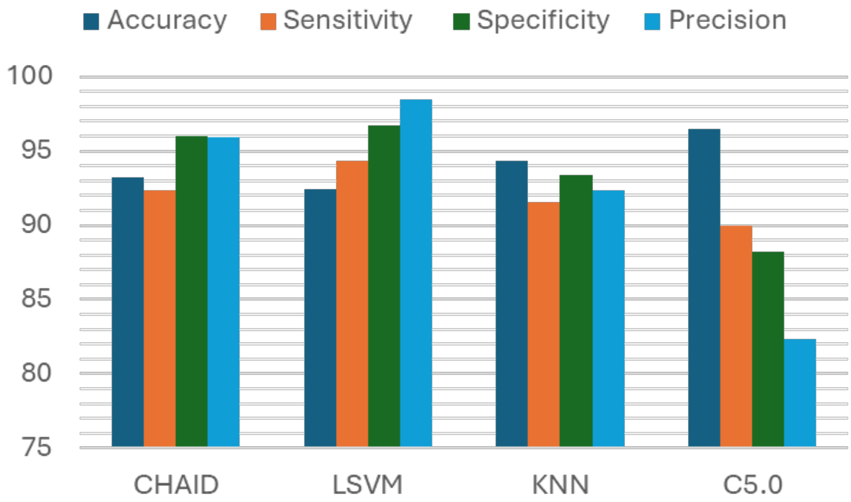

Figure-13 | Performance of top four classifiers with LASSO feature selection and SMOTE

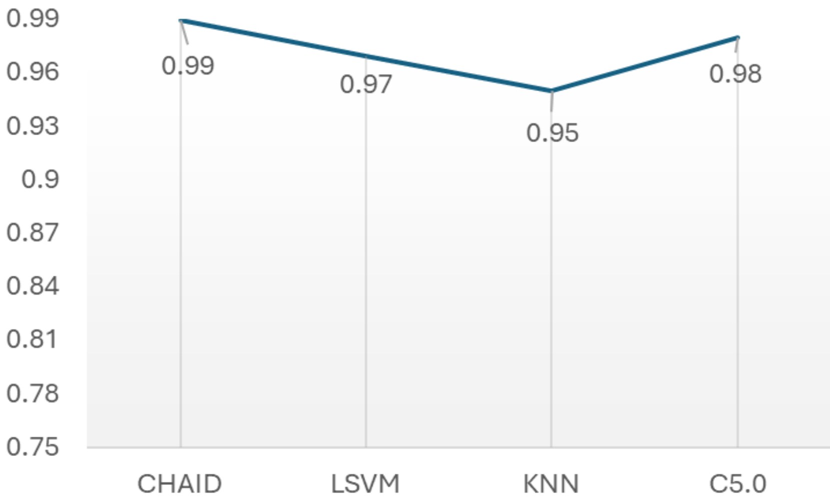

Figure-14 | AUC value comparison of top four classifiers with LASSO feature selection and SMOTE

3.6. All the features with SMOTE

When SMOTE was used on the whole feature set (without selection), LSVM again had the highest accuracy (93.46%), AUC (0.94), and F1-score (95.45%), showing that it was consistently strong across all configurations. CHAID and C5.0 were next in line, and KNN was a little less accurate but still had strong sensitivity and specificity. Table 7 represents all the results.

The comparison chart of the best 4 classifiers with all the features and SMOTE for class balancing is shown in Figure 15. The AUC is displayed in Figure 16, in which LSVM achieves the highest AUC value of 0.94.

Table-7 | Performance CHAID, LSVM, KNN and, C5.0 with all the features selection and SMOTE

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	GINI	F-Measure
CHAID	92.60	90.25	92.23	93.94	0.92	0.98	96.30
LSVM	93.46	94.54	91.46	95.22	0.94	0.97	95.45
KNN	90.54	92.58	92.24	92.16	0.87	0.91	88.21
C5.0	92.12	93.78	87.51	87.54	0.91	0.94	90.90

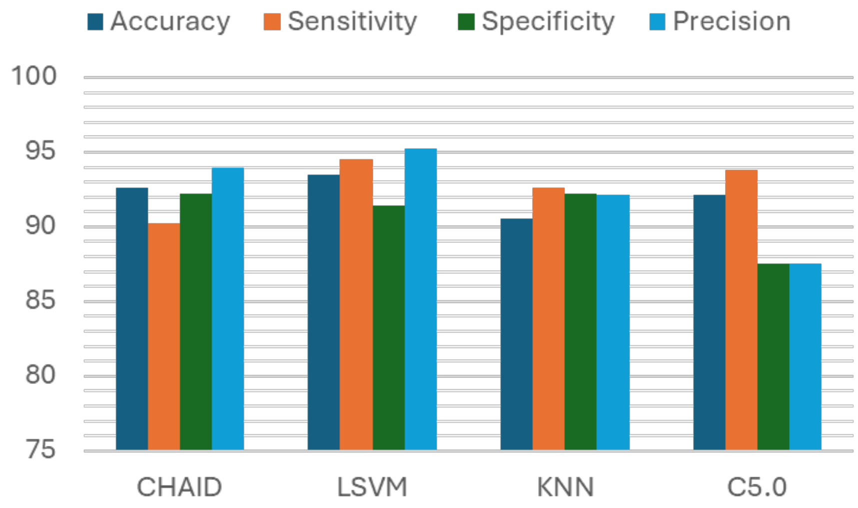

Figure-15 | Performance of top four classifiers with all the features selection and SMOTE

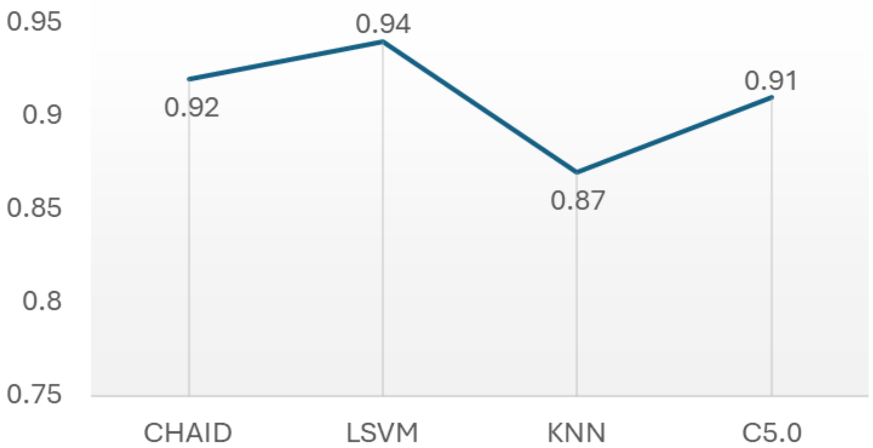

Figure-16 | AUC comparison of top four classifiers with all the features selection and SMOTE

Table-8 | Summary of the key results from the work.

Configuration	Classifier(s)	Top Accuracy	Insights
No Feature Selection	C5.0	95.23%	Better results in AUC and GINI, but sensitivity is low
CFS	LSVM	93.18%	F1-score is best among all
WFS	C5.0	96.10%	Yields the best accuracy from the setup
LASSO	CHAID	94.66%	Overall best AUC (0.96), interpretable
LASSO + SMOTE	CHAID, LSVM	93.23%, 92.42%	Excellent balance between all metrics
All Features + SMOTE	LSVM	93.46%	Stable performance, strong AUC and F1

4. CONCLUSIONS

This research introduces a machine learning approach for the early diagnosis of dementia using clinical features from the OASIS dataset. Results show that the accuracy of the diagnosis depends a lot on the appropriate feature selection method and data balancing techniques. Results show that LSVM and CHAID consistently performed well and balanced different aspects of model performance across different tests. With LASSO feature selection and SMOTE, the better balance between sensitivity, specificity, the C5.0 model with Wrapper Forward Selection (WFS) had the highest accuracy of 96.10% among all configurations. These findings imply that ensemble tree-based and margin-based classifiers are especially well-suited for structured clinical data, and they highlight the importance of customizing feature selection techniques to model architectures.

The analysis also found that a few key features from the dataset, such as MMSE score, age, and education level, remain consistent across different feature selection methods. This stability supports the fact that these features are important markers for the diagnosis of dementia at early stage.

Future work will involve integrating multimodal data (such as imaging and genomics), applying deep learning models, improving interpretability through explainable AI, and validating the framework on larger, diverse datasets.

AUTHOR CONTRIBUTIONS

Akhilesh Deep Arya: Conceptualization and methodology.

Sourabh Singh Verma: Data curation and supervision.

Prasun Chakrabarti: Formal analysis.

Rimpy Bishnoi: Validation, review and editing.

DATA AVAILABILITY

Dataset used for this research work is downloaded from Open Access Series of Imaging Studies (OASIS: https://sites.wustl.edu/oasisbrains/). It is an open access dataset and available for researchers to use for their research work.

REFERENCES

[1] Alzheimer's Association, 2022. Alzheimer's Disease Facts and Figures. Washington.

[2] Behl, T., Kaur, I., Sehgal, A., Khandige, P.S., Imran, M., Gulati, M., Anwer, M.K., Elossaily, G.M., Ali, N., Wal, P., et al., 2024. The Link Between Alzheimer’s Disease and Stroke: A Detrimental Synergism. Ageing Res. Rev., 102388.

[3] Arya, A.D., Verma, S.S., Chakarabarti, P., et al., 2023. A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer’s disease. Brain Inform., 10, 17.

[4] Dhanka, S., Sharma, A., Kumar, A. *et al.* Advancements in Hybrid Machine Learning Models for Biomedical Disease Classification Using Integration of Hyperparameter-Tuning and Feature Selection Methodologies: A Comprehensive Review. *Arch Computat Methods Eng* **33**, 289–324 (2026). https://doi.org/10.1007/s11831-025-10309-5.

[5] Marcus, M., Bernick, B.M., Moseley, S.S., Thompson, J.L., Tseng, T.M., Weinberger, D.L., 2007. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci., 19(9), 1498–1507.

[6] Bhat, S.S., Ansari, G.A. & Ansari, M.D. Performance Analysis of Machine Learning Based On Optimized Feature Selection for Type II Diabetes Mellitus. *Multimed Tools Appl* **84**, 4945–4964 (2025). https://doi.org/10.1007/s11042-024-19000-6.

[7] Song, YW., Wang, JS., Qi, YL. *et al.* Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data. *Artif Intell Rev* **58**, 111 (2025). https://doi.org/10.1007/s10462-024-11029-1.

[8] Bilal, O., Hekmat, A., Shahzad, I. *et al.* Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization. *Rev Socionetwork Strat* **19**, 271–300 (2025). https://doi.org/10.1007/s12626-025-00190-w.

[9] Gajula, S., Rajesh, V., 2024. An MRI brain tumour detection using logistic regression-based machine learning model. Int. J. Syst. Assur. Eng. Manag., 15(1), 124–134.

[10] Naresh, V.S., Reddi, S. Exploring the future of privacy-preserving heart disease prediction: a fully homomorphic encryption-driven logistic regression approach. *J Big Data* **12**, 52 (2025). https://doi.org/10.1186/s40537-025-01098-6

[11] Saputra, R.A., Agustina, C., Puspitasari, D., Ramanda, R., Warjiyono, Pribadi, D., Lisnawanty, Indriani, K., 2020. Detecting Alzheimer's Disease by The Decision Tree Methods Based On Particle Swarm Optimization. J. Phys.: Conf. Ser., 1641.

[12] Navita, Mittal, P., Sharma, Y.K. *et al.* Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease. *Int J Comput Intell Syst* **18**, 51 (2025). https://doi.org/10.1007/s44196-025-00771-1

[13] Rallabandi, V.S., Tulpule, K., Gattu, M., Initiative, A.D.N., 2020. Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer’s disease using structural MRI analysis. Inform. Med. Unlocked, 18(2), 100305.

[14] Menagadevi, M., Mangai, S., Madian, N., Thiyagarajan, D., 2023. Automated prediction system for Alzheimer detection based on deep residual autoencoder and support vector machine. Optik, 272, 170212.

[15] Houria, L., Belkhamsa, N., Cherfa, A., Cherfa, Y., 2023. Multimodal magnetic resonance imaging for Alzheimer's disease diagnosis using hybrid features extraction and ensemble support vector machines. Int. J. Imaging Syst. Technol., 33(2), 610–621.

[16] Biswas, R., Gini, R.J., 2024. Multi-class classification of Alzheimer's disease detection from 3D MRI image using ML techniques and its performance analysis. Multimed. Tools Appl., 83(11), 33527–33554.

[17] Dasu, M.V., Priyanka, C., Tejaswini, G., Ramana, J.V., Suleman, D., Tharun, S.S., 2024. Classification of Alzheimer Disease using Machine Learning

Algorithm. In: Proc. 2024 Int. Conf. Autom. Comput. (AUTOCOM), pp. 215–221.

[18] Rimal, Y., Sharma, N., Paudel, S. *et al.* Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. *Sci Rep* **15**, 13444 (2025). https://doi.org/10.1038/s41598-025-93675-1

[19] Sudharsan, S.M., Thailambal, G., 2021. Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). Mater. Today Proc.

[20] Davodabadi, A., Daneshian, B., Saati, S., Razavyan, S., 2023. Mathematical Model and Artificial Intelligence for Diagnosis of Alzheimer's Disease.

[21] Asheer, S., Bhatia, S., Sakri, S.B., 2021. Computational modeling of dementia prediction using deep neural network: Analysis on OASIS dataset. IEEE Access, 9, 42449–42462.

[22] EL-Geneedy, M., Moustafa, H.E.-D., Khalifa, F., Khater, H., AbdElhalim, E., 2023. An MRI-based deep learning approach for accurate detection of Alzheimer's disease. Alex. Eng. J., 63, 211–221.

[23] Helaly, H.A., Badawy, M., Haikal, A.Y., 2021. Deep learning approach for early detection of Alzheimer's disease. Cogn. Comput., 21(1), 1–17.

[24] Kim, J.S., Han, J.W., Bae, J.B. *et al.* Deep learning-based diagnosis of Alzheimer's disease using brain magnetic resonance images: an empirical study. *Sci Rep* **12**, 18007 (2022). https://doi.org/10.1038/s41598-022-22917-3

[25] Arya, A.D., Singh, S.V., Chakarabarti, P., Bishnoi, R., 2023. Prediction of Alzheimer's disease - A machine learning perspective with ensemble learning. In: Proc. 6th Int. Conf. Contemp. Comput. Inform. (IC3I), pp. 2308–2313. doi:10.1109/IC3I59117.2023.10397683.

[26] Srithaja, N.S., Sandhya, N., Reddy, A.B., 2023. Machine learning framework for stagewise classification of Alzheimer's disease. In: Proc. Int. Conf. Cogn. Intell. Comput. (ICCIC 2021), vol. 2, pp. 285–296. Springer.

[27] El-Gamal, F., Elmogy, M., Mahmoud, A., Shalaby, A., Switala, A.E., et al., 2021. A personalized computer-aided diagnosis system for mild cognitive impairment (MCI) using structural MRI (sMRI). Sensors, 21(16), 5416.

[28] assawab, R., Ouzir, M., Benyacoub, B., El Allati, A., & El Moudden, I. (2025). Medical image feature extraction and selection based on InceptionV3 and Gini Index for cervical cancer cells identification. *Statistics, Optimization & Information Computing*, *15*(2), 1050-1067. https://doi.org/10.19139/soic-2310-5070-2525

[29] Li, J., 2024. Area under the ROC Curve has the most consistent evaluation for binary classification. PLOS ONE, 19(12), 1–28.