Sigma J Eng Nat Sci, Vol. 43, No. 5, pp. 1484−1494, October, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00147

Research Article

Cart KNN: A hybrid distance-based pre-pruned classification and regression method for the early diagnosis of Alzheimer's disease

Pijush DUTTA[1,*], Sunita SHARMA[2], Varun Kumar NOMULA[3], Gour Gopal JANA[1], Anubrata MONDAL[4], Shobhandeb PAUL[5]

[1] *Department of Electronics and Communication Engineering, Greater Kolkata College of Engineering and Management, West Bengal, 743387, India*
[2] *School of Biotechnology, University Jawaharlal Nehru University, New Delhi, 110067, Delhi*
[3] *Principal AI/ML Engineer, Apex IT Services LLC, Atlanta, 10632, USA*
[4] *Department of Electrical Engineering, Greater Kolkata College of Engineering and Management, West Bengal, 743387, India*
[5] *Data Scientist, TCS Oval, Newtown, West Bengal, 700160, India*

ARTICLE INFO

Article history
Received: 01 April 2024
Revised: 12 June 2024
Accepted: 24 September 2024

Keywords:
Alzheimer's Disease Prediction; CART-KNN; Confusion Matrix; Decision Tree; K- Nearest Neighbors; Pruning

ABSTRACT

Alzheimer's disease is a primary cause of dementia in the aged population furthermore, a significant segment of the global populace experiences metabolic disorders including diabetes and Alzheimer's disease. Alzheimer's disease has a damaging effect on the brain. Due to its detrimental effects on memory and physical functioning, this condition may lead to an increase in the number of inactive individuals as the senior population rises. To identify these diseases early on, researchers have looked into a variety of deep learning and machine learning techniques. Patients with Alzheimer's disease (AD) can recover from it more successfully and with less damage if they receive early diagnosis and therapy. Decision trees are used in conjunction with K-nearest neighbor and classification and regression trees to forecast and recommend the AD. The overall experiment is represented in three stages, in the first stage a comparative study was analyzed with the Proposed model along with its fundamental algorithm k-nearest neighbor algorithm (KNN) and detection threshold (DT), and in the second stage four performance indicators of the confusion matrix were used to demonstrate that our suggested strategy performs better in terms of running time performance metrics, accuracy, precision, and F1 score. In the final stage, the present experiment demonstrates a notable improvement in running time over the conventional k-nearest neighbor technique with an accuracy of 98.23%, the suggested model outperforms other cutting-edge methods for the OASIS dataset. Moreover, the proposed model also has better Moreover experimental result also suggests that the proposed CART-KNN model performed better than other general models: K-nearest neighbor and Decision tree by showing smaller root mean square error (RMSE) and mean absolute percentage error (MAPE).

Cite this article as: Dutta P, Sharma S, Nomula VK, Jana GG, Mondal A, Paul S. Cart KNN: A hybrid distance-based pre-pruned classification and regression method for the early diagnosis of Alzheimer's disease. Sigma J Eng Nat Sci 2025;43(5):1484−1494.

***Corresponding author.**
*E-mail address: pijushdutta009@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey

INTRODUCTION

The most prevalent type of dementia that needs intensive medical attention is Alzheimer's disease (AD). An accurate and timely study of AD prognosis is required to start clinical progress and provide effective patient therapy [1]. AD is a long-term, degenerative brain disease that gradually destroys brain tissue, and impairs thinking and memory [2]. The amount of computing power used by healthcare departments is always growing, and patient data is increased electronically. Although many electronic health records (EHRs) are now easier to access, 80% of the data remains unstructured. Because of this, handling unstructured data with database management software and other conventional techniques is difficult. Some methodologies adopted for EHRs through machine learning and data mining technologies improve the standard and efficiency of healthcare and prescribe proper medicine [3]. Finding undiscovered and meaningful patterns from a large number of pre-existing datasets is a process known as data mining or knowledge discovery. These patterns aid in the comprehension of the historical dataset, the classification of a new set of data, and the creation of data summaries. Data mining categorizes or groups records according to their similarities or differences, hence aiding in identifying deeper patterns within the data[4]. Data mining has been widely applied in various fields over the past few decades, including marketing, retail, finance, stock market forecasting, and medical and healthcare [5]. Machine learning and data mining techniques have been widely employed in medical and healthcare research [6]. These algorithms may be used to categorize distinct subjects based on the similarities in their qualities. The primary goal of this research project is to identify the various phases of Alzheimer's disease (AD) using machine learning and data mining techniques on the Open Access Series of Imaging Studies (OASIS). Finding the most distinctive feature for each of the several phases of AD within the OASIS dataset is also a sub-objective of this research project. Several studies focus on using machine learning to identify blood-based non-amyloid biomarkers for early Alzheimer's disease detection, achieving high sensitivity and specificity levels with novel protein panels [7]. Machine learning, particularly the Random Forest model, achieves 90.6% accuracy in early Alzheimer's detection using neuroimaging biomarkers, emphasizing its robustness and diagnostic advancements in Alzheimer's disease prediction [8]. Random Forest, XGBoost, and Light GBM were used to predict Alzheimer's disease risk in individuals aged 65+ using Finnish register data, showing low predictive accuracies and suggesting additional data sources were needed [9]. Gradient boosting, random forest, and support vector machines utilizing EEG signals and genetic information effectively classifies Alzheimer's disease, achieving high accuracy and offering a comprehensive diagnostic approach beyond imaging limitations [10]. Novel drug discovery for Alzheimer's disease utilized machine learning to identify compounds targeting key proteins like MTOR and BCL2, offering a multi-targeted therapeutic approach for AD treatment [11]. An explainable machine learning approach for Alzheimer's disease classification achieved high accuracy using SVM models and highlighted key factors like MEMORY and JUDGMENT in AD development [12]. Machine learning-based speech analysis can effectively distinguish between different cognitive impairments in the Alzheimer's disease spectrum and predict cognitive domain performance, offering the potential for early detection and monitoring [13].

In recent research data modeling is quite challenging and complex during big data and short-term prediction. A fusion of CART and KNN models has been proposed to solve this issue. A nonlinear CART model is effective for data porting and mining purposes [14]. KNN, a popular nonparametric regression model for searching similar data in a particular region [15]. A novel hybrid approach of CART and KNN has been experimented with for software fault prediction. The proposed approach achieved an average accuracy of 0.90274 which is much higher than the other nine machine-learning models [16]. This work's main contributions are: to consider the input feature and perform the autocorrelation for finding out the data pattern similarity by CART model. Secondly, combined a regression tree with weighted k-nearest neighbors and distance-based pre-pruned classification to create a hybrid technique.

The present research is arranged into the following stages: in the literature review section, Research gaps and earlier studies are described. The proposed methodology including dataset description, preprocessing, pruning, and CART-KNN method explained. The final section experimentation and performance assessment of our suggested methodology are covered.

Literature Review

Feature selection and extraction techniques are used in machine learning (ML) algorithms to predict Alzheimer's disease. The Oasis dataset is used as the basis for the classification process. A quick summary of the various methods used in brain image analysis for brain disease diagnosis is given [17]. Based on the findings, this article discusses several important brain disease diagnoses. This work intends to integrate contemporary research on Alzheimer's disease using different AI platforms. The most accurate diagnosis approach may be identified by the authors by utilizing 21 brain illness reviews. Deep convolutional autoencoders investigate AD data analysis where the proposed method extracts MRI characteristics that reflect neurodegeneration and cognitive symptoms from MRI images [18].With imaging-derived indicators, MMSE or ADAS11 scores can be utilized with over 80% accuracy to predict the diagnosis of AD. To achieve binary classification, a deep neural network with linked layers is employed [19, 20] and to activate each concealed layer distinct activation function was used. A lack of education, high blood pressure, obesity, hearing loss, depression, diabetes, inactivity, smoking, and social isolation are some of the variables that might increase one's risk.

Deep Learning with CNN methods has been proposed to extract MRI features for early Alzheimer's diagnosis, achieving over 97% accuracy, and enhancing detection in its earliest stages [21]. ML models like Gaussian NB, Decision Tree, and Voting Classifier can predict AD disease early with high accuracy, aiding in successful treatment and minimizing harm. Machine learning algorithms can accurately predict Alzheimer's disease with a 96% validation accuracy[22]. A CNN-SVM model was investigated for early Alzheimer's disease prediction using MRI images, achieving 94.44% accuracy on Mild Cognitive Impairment (MCI) subjects [23]. A novel ensemble machine learning approach, incorporating feature selection and adaptive voting, was examined in early Alzheimer's disease detection. The proposed model achieved 93.92% accuracy outperforming traditional methods by 3.39% [24]. A classification model was proposed using SVM, RF, and FNN algorithms for early detection of Alzheimer's disease, enhancing efficiency and reducing processing time in medical data analysis [25]. A speech-based machine learning has been designed for early detection of Alzheimer's disease with a 75.59% accuracy, showing promise for non-invasive diagnosis [26]. Machine learning models like DT, RF, and SVM have been investigated to enable early Alzheimer's disease prediction with up to 80% accuracy using OASIS data [27]. Machine learning models, like Gaussian Naive Bayes, have been proposed to accurately predict Alzheimer's disease early by analyzing neuroimaging and cognitive data with a classification accuracy of 96.92% [28]. Machine learning models like DT, RF, SVM, Voting classifiers, and GB have been investigated for the Prediction of early stages of Alzheimer's disease, achieving 83% maximum accuracy on test data [29]. A novel algorithm has been developed using a modified capsule network on the OASIS dataset, achieving 92.39% accuracy in predicting dementia, and outperforming traditional methods and deep learning classifiers [30].

After reviewing earlier research, we created a hybrid strategy based on decision trees and k-nearest neighbors. The following is a list of the proposed work's primary contributions:

1. a novel method known as distance-based pruning is suggested to prune decision tree nodes. The last part provides a detailed explanation of the distance-based pruning approach's processes.
2. Our suggested method reduces the O(n) running cost of the standard k closest neighbor technique to O (log n) + c explained in section 3. KNNs are introduced to the decision tree's leaf nodes to lower running costs during the training phase.
3. The CART is created using decision tree's-based leaf nodes, k closest neighbors are placed in place of class labels.
4. To improve the effectiveness of conventional k nearest neighbors, the idea of weights is added in the prediction phase, based on the sigmoid function. The closest neighbors' distance from the site under consideration inversely determines the weights.

WORKING APPROACH

This paper suggests a hybrid method based on weighted k-nearest neighbors, regression trees, and distance-based pre-pruned categorization. The hybrid CART-KNN approach aims to improve prediction accuracy and generalization by combining the decision tree's structured, rule-based model with KNN's instance-based, flexible learning. There are various ways to hybridize CART and KNN, but a common approach involves using CART to preprocess data or make initial predictions and then applying KNN to refine these predictions. The overall process can be performed in three stages: tree-based segmented preprocessing, KNN-based post-processing, and finally prediction of the hybrid model. During the Tree-Based Segmented preprocessing stage CART is used to partition the feature space into distinct regions based on the decision tree's splits. Now KNN algorithm will only consider neighbors within the same region, potentially leading to more localized and accurate predictions. In KNN as a Post-Processing Step after that KNN refines and boosts lower tree node confidence of CART's prediction. During the Hybridization of the Predicted Model, the prediction can be done by either combining predictions by using weighted average prediction obtained by CART and KNN or a voting mechanism where the best prediction is considered by a majority vote from both models. This methodology helps Hybrid CART-KNN to reduce the Overfitting of the distribution of data, improve Locality Sensitivity by handling varying data densities, and flexibility for adapting to different types of data and problems. The run time of the proposed algorithm is reduced to O (log n) +c from O(n), where each factor has its significance like the notion of O(n) is linear time, while O (log n) running times divide-and-conquer application because of ideally cutting the work in half every time. m-dimensional points represent independent characteristics in the dataset. A K – Matrix is generated with the following condition; the element $e_{ij} = 0$ if the jth training sample is far from the ith training sample otherwise $e_{ij} = 1$. Following the KNN-matrix computation, the Euclidean distance formula is used to determine the maximum distance Max distance between all points, and a constant parameter called tolerance is added to regulate the creation of decision trees. Each node of DT fulfills the criteria $Max_{distance} *$ tolerance, then the KNN matrix stores the leaf node value rather than class labels. The Weighted K-Tree (WK Tree) creation technique is described in full in this portion of the study publication [31]. These are the steps that explain how to construct a WK-Tree:

Pseudo Code

Let the Input contain, Training Samples X, and Testing Classes Y, and the Output contain Confusion Matrix

Step 1: m dimensional space is considered for all training samples and Euclidean distance is measured by eqn. (1).

Step 2: KNNs construct a matrix by considering all training samples using eqn. (2).

Step 3: Applied CART, a distance-based pre-pruning

Step 4: KNN stores all the leaf nodes without repetition and removes all duplicate data.

Step 5: The decision Tree helps the testing samples to try to reach the leaf node.

Step 6: The weighted label of each test point is assigned.

Step 7: Based on predicted and actual values Confusion matrix is designed.

Step 8: Accuracy, sensitivity, Precision, and F1-Score are calculated with the help of Confusion Matrix

Calculation of the Largest Distance

The approach's initial step involves treating all training samples lying in m-dimensional space and computing the maximum distance between them all. Distance in m dimensions was calculated between all training samples [32]. Equation (1) may be used to determine the Euclidean distance or space,

$$d(P_i,P_j) = max_{\ 1\leq i,j\leq n} \sqrt{\sum_{k=1}^{m} (P_i^k - P_j^k)^2} \qquad (1)$$

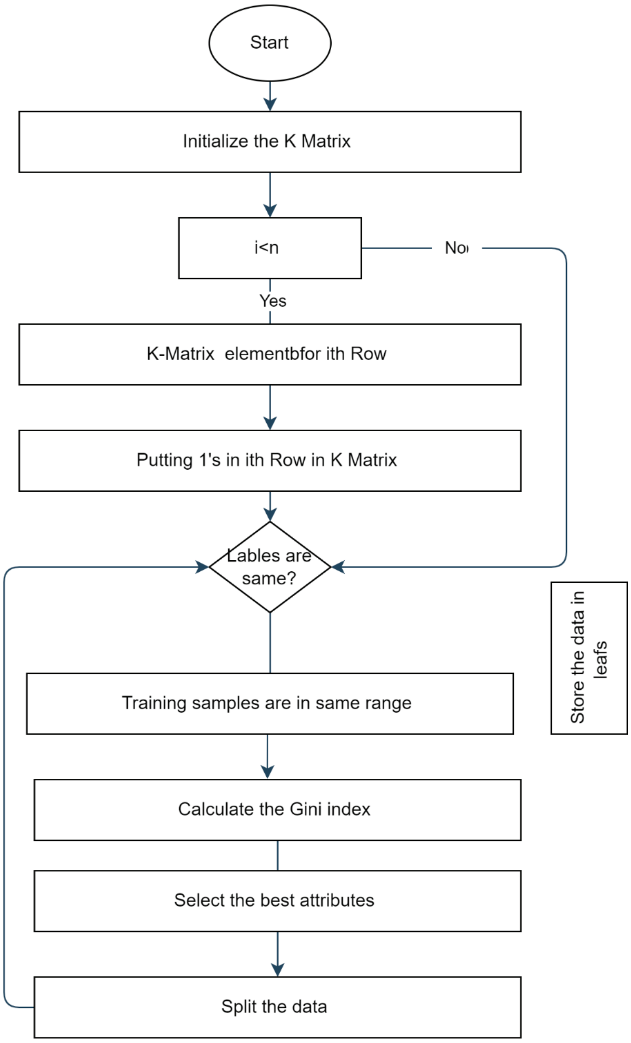
Start
Initialize the K Matrix
i<n
No
Yes
K-Matrix elementbfor ith Row
Putting 1's in ith Row in K Matrix
Lables are same?
Store the data in leafs
Training samples are in same range
Calculate the Gini index
Select the best attributes
Split the data

Figure 1. Flow Diagram of the present research.

Iterating over the number of dimensions is the variable k while iterating over the number of training samples is the variable i & j.

KNN Matrix Generation

For every training sample, a n ∗ n matrix of KNN is constructed. The function $\sqrt{n}$ + c is chosen to determine the closest neighbors. Equation (2) displays an example of a 5 ∗ 5 KNN matrix, where all diagonal components are equal to 1. Position j is regarded as the closest neighbor of position i if the element e_{ij} of the K- -matrix is 1, but not the nearest neighbor point if the element e_{ij} of the KNN matrix is 0 (Fig. 1).

$$\text{KNN-Matrix} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix} \qquad (2)$$

Decision Tree Creation

Using distance-based pre-pruning and the Gini index node splitting approach, a decision tree is constructed [33]. Pre- and post-pruning are the two types of DT node pruning techniques. Since pre-pruning based on each leaf node's maximum depth from the tree's root is ineffective, we have instead used pre-pruning based on distance. Distance-based pruning involves multiplying the maximum distance by the constant parameter Tolerance, which is initially set between 0 and 1.

Labeling of Testing Samples

During the labeling step, the testing sample's final label is assigned by applying the weighted k-nearest neighbor technique after traversing the classification and regression trees to reach the leaf node [34]. Using equation (3) weights of each neighbor were calculated.

$$w_i = (1.0 - \frac{1}{1+ e^{-d_{ij}}}) *L_i \qquad (3)$$

Where L_i indicates fault and non-fault-prone classes.

Distance-Based Pruning

The greatest distance between any two points is determined first in distance-based pre-pruning of decision trees, and the decision tree is subsequently pruned using the percentage of the maximum distance. To encompass all locations, the maximum distance is computed [35]. Max distance and the global parameter are computed. Max distance ∗ Tolerance is checked for all nodes, and tolerance is manually set between 1 and 0 according to the point density. Expression of Parameter Max distance is.

Time Complexity

Since training time is a one-time expense, our suggested approach's testing time complexity will be the only thing we talk about [36]. During the prediction phase, the Decision

Tree traverses in O (log n) time complexity from the root node of each leaf. In our suggested technique, k-nearest neighbors are kept at leaf nodes instead of labels, adding a little constant c to the decision tree's real testing time complexity. Our suggested method offered total run time complexity is O (log n + c).

Data Preprocessing

There are missing values and redundant data in the raw dataset [37-40]. Features with missing values are extracted and modified as part of data management. In this study, min-max feature scaling methods proposed to scale the feature between the range (0,1) with the help of minimum value & difference of the range and information gain filtering methods used for feature selection carried out in preparation for the KNN algorithm's preprocessing of the dataset.

Data Analysis

Examining the dataset, followed by cleaning, converting, and modeling it into a format that works, is data analysis. Finding useful information that will subsequently be utilized to enhance decision-making is the goal of data analysis [41,42]. The study might be useful, for instance, in assessing the properties of the data and the relationship between co-relation features.

Dataset

This study uses a dataset from the Open Access Series of Imaging Studies (OASIS) [27] to predict Alzheimer's disease. Depending upon cross-sectional MRI and longitudinal MRI the data samples are assigned as dementia and non-dementia at a certain time or baseline is the first step. 150 people, ages 60 to 96, whose MRI data are included in the study. The dataset contains 72 non-demented patients and 69 were dementia. This distinction did not change during the study. The twelve OASIS longitudinal dataset characteristics are displayed in Table 1. This includes handling missing values, extracting and transforming features, and so on. Nine rows in the SES column have missing values, which may be fixed by removing the corresponding rows [43,44]. The measurement as a whole is now 141.

Table 1. List of twelve attributes

Identification, Age, Years of Education, Gender, Dominant Hand, Socio-Economic Status (SES), Clinical Dementia Rating, Normalize Whole Brain Volume, Mini-mental state examination, Atlas Scaling Factor, Estimated Total Intracranial Volume, Delay.

RESULTS AND DISCUSSION

This article's suggested method was created and tested on a computer with an 8GB RAM corei5 CPU. This method is developed using Anaconda 3, and its performance is evaluated against various machine learning models [45,46]. Before we proceed to the comparative study of proposed methods with the other three machine learning algorithms we find the effectiveness of the CART-KNN with the other two fundamental machine learning algorithms KNN and DT in Table 2. Cohen Kappa ranking is used to determine the degree of agreement between two independent raters and AUC (Area under the curve) for binary classification of the dataset are the main metrics to evaluate.

Accuracy

This study employs a suggested methodology to https://www.kaggle.com/jboysen/mri-andalzheimers?select=oasis_cross-sectional.csv is where you can find this data. Table 3 presents an accuracy performance metric comparison between our suggested method and three SVM, RF, CART-KNN, and decision tree variants (CART). To complete the experiment, set the tolerance value to 0.05 [47].

The accuracy of our suggested method is compared to various machine-learning techniques in Figure 2. The accuracy distribution of our suggested method demonstrates unequivocally that it performs better in this article than alternative machine learning techniques [48].

Table 2. Comparison study of the proposed approach and its fundamental approaches

Methods	Choen Kappa	AUC	Accuracy
KNN	0.887	0.891	90.85%
DT	0.812	0.837	82.55
CART-KNN	0.965	0.98	98.23%

Table 3. A comparative study based on accuracy for constant tolerance =0.05

s	DT	RF	SVM	CART-KNN
Alzheimer's dataset	82.55%	89.64%	84.72%	98.23%
CART:				

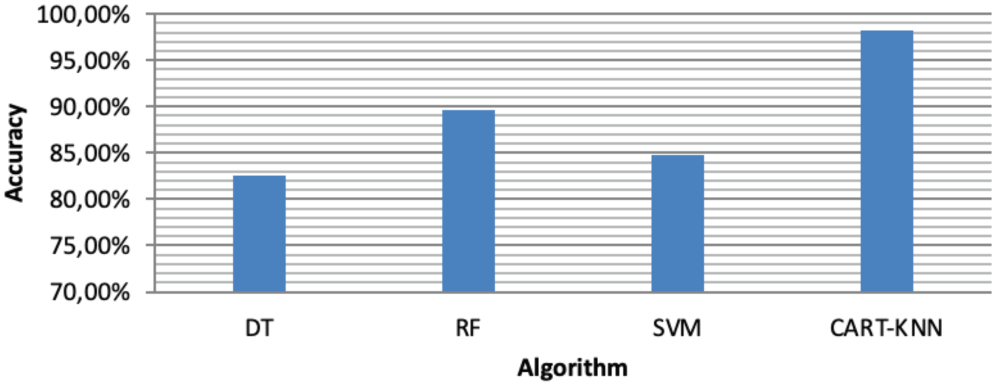

Figure 2. Comparison of accuracy in terms of bar plots.

Other Performance Metrics

The average accuracy, F1-score, and recall, for all methods compared on datasets related to Alzheimer's disease are displayed in Table 4. This article's suggested strategy outperforms all other approaches utilized in the comparison in terms of accuracy and f1-score performance criteria. A graphical depiction of the accuracy, recall, and F1-score comparison research is shown in Figure 3.

Run Time

The run time of the proposed model is O (log n) + c, which is compared with another traditional model in Figure 4.

Table 5 presents a comparison with previous research on the OASIS dataset for AD prediction using several machine-learning algorithms. Kavitha and others. [29] demonstrated 83% accuracy in AD prediction using ML on the OASIS dataset, Basheer et al. Using a combination of deep learning and machine learning, [30] achieved a model accuracy of around 92.39%. Six distinct ML approaches were developed by Uddin et al. [22] and the developers of one of them achieved a maximum accuracy of 96% for the voting algorithm.

Comparison of Prediction Accuracy of the Model

In this part, we assessed the two types of error metrics, MAPE and RMSE. RMSE quantifies the error's variability

Table 4. Comparison of Precision, Recall, and F1-score of the proposed method with other approaches

Techniques	Precision	Recall	F1-score
DT	0.8108	0.7956	0.7928
RF	0.863	0.852	0.851
SVM	0.832	0.827	0.82
CART-KNN	0.972	0.969	0.965

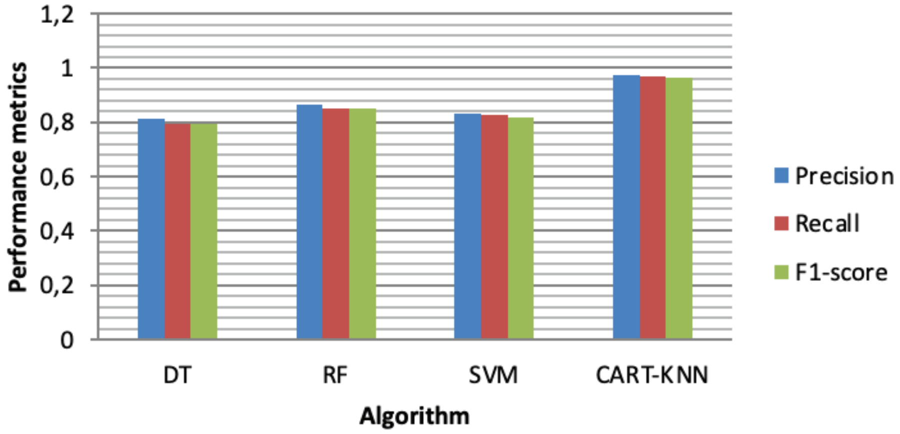

Figure 3. Graphical comparison for other performance metrics.

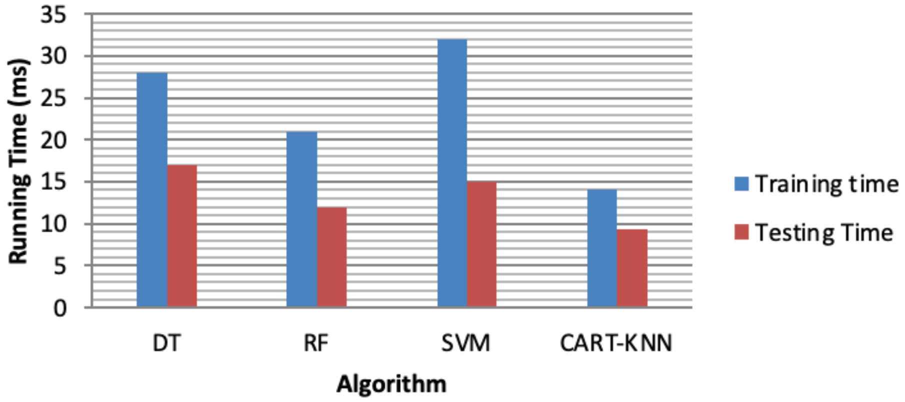

Figure 4. Comparison of Computational time.

Table 5. Comparison with state of art methods based on the OASIS dataset

Reference	Techniques	Outcomes of the algorithm
Kavitha et al. [29]	Gradient Boosting, SVM, DT and Voting	Validation accuracy 83%
Basheer et al.[30]	Machine learning and deep learning models(M-CapNet)	acceptable accuracy of 92.39%
Uddin et al. [22]	Gaussian NB, DT, RF, XGBoost, Voting Classifier, and Gradient Boost	Maximum accuracy obtained 96%
Proposed model	CART- KNN model	Acceptable accuracy about to 98.23%

while MAPE calculates the difference between the predicted and the actual result [49,50]. The following equation can express the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{N}} \quad (4)$$

$$MAPE = \frac{\sum_{i=1}^{N}\left|\frac{y_i - \bar{y}_i}{\bar{y}_i}\right|}{N} * 100 \quad (5)$$

where N, y_i and $\bar{y}_i$ are several samples, actual and predicted output respectively. A lower value of MAPE and RMSE indicates a better prediction. Figure 5 and Figure 6 represent the optimum parametric condition of the proposed CART -KNN model.

Table 6. Computational value of F1 score for CART-KNN model for different leaf nodes value

Leaf nodes	Testing F1 score	Training F1 Score
0	0.26	0.265
10	0.43	0.46
20	0.47	0.52
30	0.48	0.53
40	0.48	0.54
50	0.48	0.55

The F1 score for the train and testing dataset almost ranges the same lies from 0.25 to 0.54 for different values of no. of the leaf node shown in Table 6, while MAPE is increased for the training dataset by increasing the number of neighbors, but for the testing dataset, MAPE is reversed shown in Table 7.

Table 7. Computational value of MAPE for CART-KNN model for different neighbor value

No of Neighbour	Testing Data	Training Data
0	0.32	0.14
5	0.28	0.14
10	0.25	0.15
15	0.25	0.16
20	0.24	0.19
25	0.23	0.22

Table 8. Comparative study on RMSE and MAPE

Metrics	Algorithm		
	KNN	DT	CART-KNN
RMSE	205278	198634	50678
MAPE	14.98	10.98	8.56

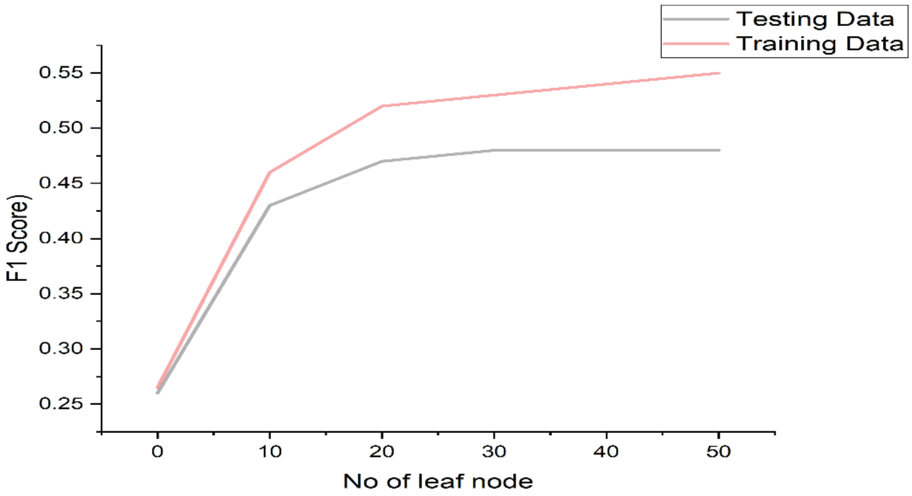

Figure 5. F1 Score for CART-KNN Cross-validation selection.

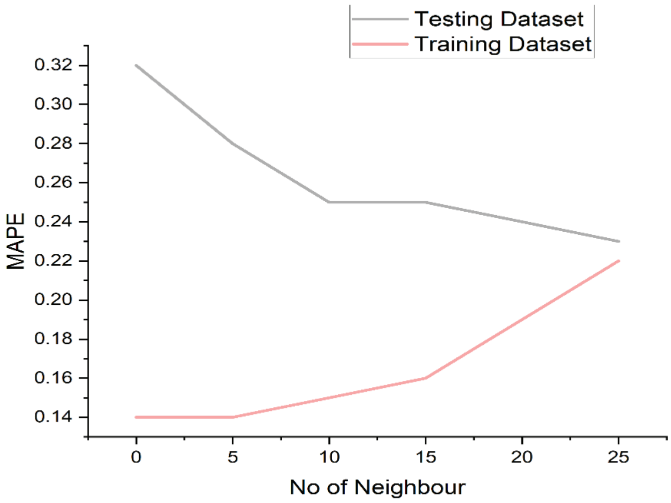

Figure 6. MAPE for CART-KNN Cross-validation selection.

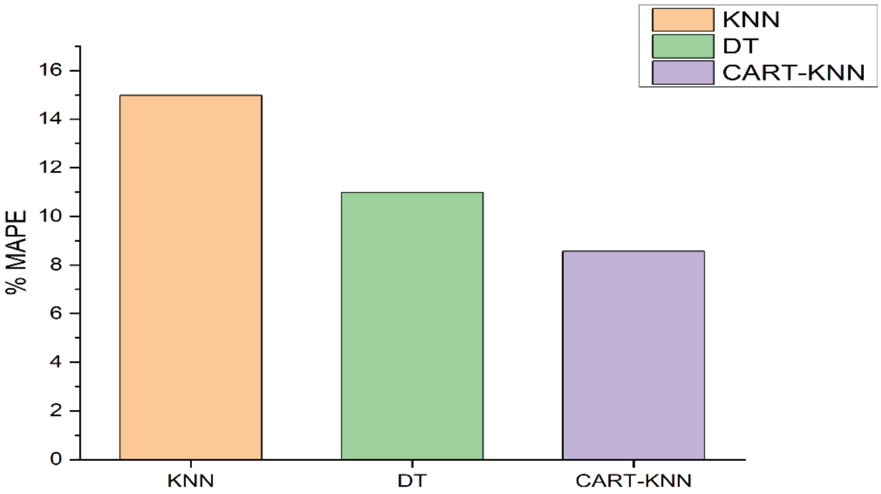

Figure 7. Comparison of % MAPE for all model.

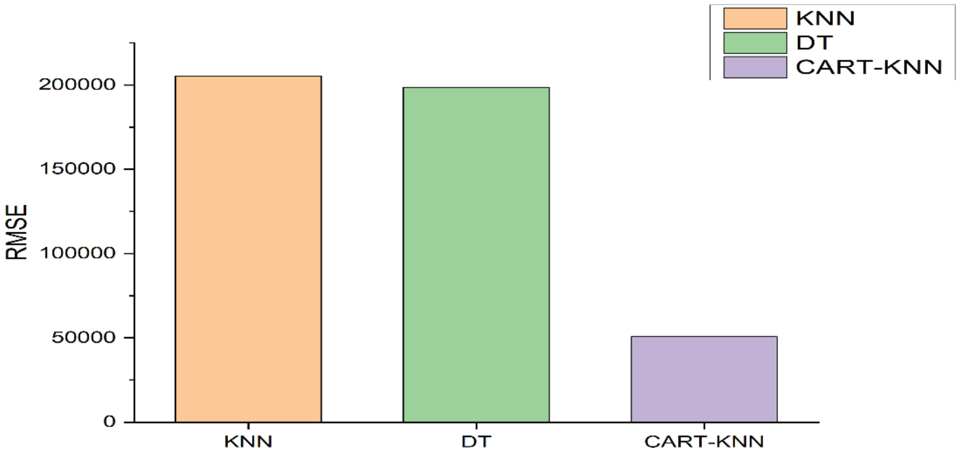

Figure 8. Comparison of RMSE for all model.

Figure 7 and Figure 8 show the prediction performance of the proposed hybrid model and traditional models. The proposed fused CART-KNN model improved the RMSE by 75.31%, and 74.31% while MAPE by 42.85%, and 2.20% respectively concerning traditional DT, and KNN models. Furthermore, Table 8 shows that the hybrid model is closest to the real value and has minimal swings in prediction errors.

CONCLUSION

The hybrid strategy of K nearest neighbor and pre-pruned classification and regression tree (CART) is suggested in this article. To overcome the overfitting problem, a pre-pruned Decision tree is applied in m-dimensional space and stores training samples of leaf nodes. Open Access Series of Imaging Studies (OASIS) datasets dataset for Alzheimer's disease (AD) prediction have been used to test the suggested strategy, and the outcomes are compared with those of other cutting-edge methods based on OASIS datasets and machine learning models. The main conclusions are listed as follows:

1. The robustness of the CART-KNN model allows it to choose parameters based on the features of the time series.
2. The proposed fused CART-KNN model improved the RMSE by 75.31%, and 74.31% while MAPE by 42.85%, and 2.20% respectively concerning traditional DT, and KNN models.
3. The validation accuracy of the proposed CART-KNN is about 98.23% on the test data of AD, which indicates the CART-KNN classifier in the suggested study produces a more advantageous result.

However, this study has also some limitations: The CART-KNN suffered in poor generalization ability due to easy overfit. New hybrid approaches based on conventional machine learning techniques can be created in the future to increase performance and make them suitable for real-world applications. To make the concept feasible, work on running time complexity reduction can be done

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://www.kaggle.com/jboysen/mri-andalzheimers?select=oasis_cross-sectional.csv.

NOMENCLATURE

$d(P_i,P_j)$	Distance between two points
L_i	fault-prone classes
w_i	Weighted value value for ith neighbor
N	no. of samples
y_i	the output of ith number of samples
$\overline{y_\imath}$	Predicted output of ith samples

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Murugan S, Venkatesan C, Sumithra MG, Gao X-Z, Ekalliya B, Akila M, et al. DEMNET: A deep learning model for early diagnosis of Alzheimer diseases and dementia from MR images. IEEE Access 2021;9:90319–90329. [CrossRef]

[2] Akiyama H, Barger S, Barnum S, Bradt B, Bauer J, Cole GM, et al. Inflammation and Alzheimer's disease. Neurobiol Aging 2000;21:383–421. [CrossRef]

[3] Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A. Classification of Alzheimer's disease using machine learning techniques. Data 2019:296–303. [CrossRef]

[4] Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012;36:2431–2448. [CrossRef]

[5] Gupta MK, Chandra P. A comprehensive survey of data mining. Int J Inf Technol 2020;12:1243–1257. [CrossRef]

[6] Berkhin P. A survey of clustering data mining techniques. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional data. Berlin: Springer-Verlag; 2006. p. 25–71. [CrossRef]

[7] Dennis E, Manikandan G, Vilma V, Hemalatha S. Alzheimer disease using machine learning. Int J Sci Res Sci Technol 2025;58:262. [CrossRef]

[8] Sowjanya V, Neeha Y, Dishasri L, Neela S, Prasanth Y. Predictive diagnosis of Alzheimer's disease using machine learning. 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), 09-11 May 2024.

[9] Hao L, Sirpa H, Julian L, Vesa T, Anna-Maija T. Predicting Alzheimer's disease from cognitive footprints in mid and late life: how much can register data and machine learning help? Int J Med Inform 2024;190:105540. [CrossRef]

[10] Wujia Y, Ting-Hsuan S, Kai-Cheng H, et al. Comparative analysis of machine learning algorithms for Alzheimer's disease classification using EEG signals and genetic information. Comput Biol Med 2024;176:108621. [CrossRef]

[11] Fahad M, Alshabrmi F, Aba F, Alkhayl F. Novel drug discovery: advancing Alzheimer's therapy through machine learning and network pharmacology. Eur J Pharmacol 2024;976:176661. [CrossRef]

[12] Alatrany A, Khan W, Hussain A, Hoshang K, Al-Jumeily D. An explainable machine learning approach for Alzheimer's disease classification. Dent Sci Rep 2024;14:2637. [CrossRef]

[13] García-Gutiérrez F, Alegret M, Marquié M, Muñoz N, Ortega G, Cano A, et al. Unveiling the sound of the cognitive status: Machine Learning-based speech analysis in the Alzheimer's disease spectrum. Alzheimers Res Ther 2024;16:26. [CrossRef]

[14] Breiman L. Classification and regression trees. Abingdon: Routledge; 2017. [CrossRef]

[15] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13:21–27. [CrossRef]

[16] Singh M, Chhabra JK. A hybrid approach based on k-nearest neighbors and decision tree for software fault prediction. Kuwait J Sci 2023;50.

[17] Khan P, Kader F, Islam SMR, Rahman AB, Kamal S, Toha MU, et al. Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. IEEE Access 2021;9:37622–37655. [CrossRef]

[18] Martinez-Murcia FJ, Ortiz A, Ramirez J, Castillo-Barnes D. Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J Biomed Health Inform 2019;24:17–26. [CrossRef]

[19] Prajapati R, Khatri U, Kwon GR. An efficient deep neural network binary classifier for Alzheimer's disease classification. In: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Piscataway (NJ): IEEE; 2021. p. 231–234. [CrossRef]

[20] Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. Cogn Comput 2022;14:1711–1727. [CrossRef]

[21] Kadhim KA, Sakran AA, Adnan MM, Salman GA. Early diagnosis of Alzheimer's disease using convolutional neural network-based MRI. Malays J Fundam Appl Sci 2023;19:362–368. [CrossRef]

[22] Uddin KMM, Alam MJ, Jannat JA, Aryal S. A novel approach utilizing machine learning for the early diagnosis of Alzheimer's disease. Biomed Mater Devices 2023;1:882–898. [CrossRef]

[23] Rabeh AB, Benzarti F, Amiri H. CNN-SVM for prediction Alzheimer disease in early step. In: 2023 International Conference on Control, Automation and Diagnosis (ICCAD). Piscataway (NJ): IEEE; 2023. p. 1–6. [CrossRef]

[24] Irfan M, Shahrestani S, Elkhodr M. Early detection of Alzheimer's disease using cognitive features: a voting-based ensemble machine learning approach. IEEE Eng Manag Rev 2022;51:16–25. [CrossRef]

[25] Varma BSS, Kalyani G, Asish K, Bai MI. Early detection of Alzheimer's disease using SVM, random forest and FNN algorithms. In: 2023 2nd International Conference for Innovation in Technology (INOCON). Piscataway (NJ): IEEE; 2023. p. 1–6. [CrossRef]

[26] Tripathi T, Kumar R. Speech-based detection of multiclass Alzheimer's disease classification using machine learning. Int J Data Sci Anal 2024;18:83–96. [CrossRef]

[27] Shrivastava RK, Singh SP, Kaur G. Machine learning models for Alzheimer's disease detection using OASIS data. In: Koundal D, Jain DK, Guo Y, Ashour AS, Zaguia A, editors. Data analysis for neurodegenerative disorders. Singapore: Springer Nature; 2023. p. 111–126. [CrossRef]

[28] Mahjabeen A, Mia MR, Shariful FNU, Mahmud I. Early prediction and analysis of DTI and MRI-based Alzheimer's disease through machine learning techniques. In: Kaiser MS, Waheed S, Bandyopadhyay A, Mahmud M, Ray K, editors. Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering. Singapore: Springer Nature; 2023. p. 3–13. [CrossRef]

[29] Kavitha C, Mani V, Srividhya SR, Khalaf OI, Tavera Romero CA. Early-stage Alzheimer's disease prediction using machine learning models. Front Public Health 2022;10:853294. [CrossRef]

[30] Basheer S, Bhatia S, Sakri SB. Computational modeling of dementia prediction using deep neural network: analysis on OASIS dataset. IEEE Access 2021;9:42449–42462. [CrossRef]

[31] Banga M, Bansal A. Proposed software faults detection using a hybrid approach. Secur Priv 2023;6:e103. [CrossRef]

[32] Maddipati SS, Srinivas M. A hybrid approach for cost effective prediction of software defects. Int J Adv Comput Sci Appl 2021;12. [CrossRef]

[33] Priyanka NA, Kumar D. Decision tree classifier: a detailed survey. Int J Inf Decis Sci 2020;12:246. [CrossRef]

[34] Chen P, Liao BB, Chen G, Zhang S. Understanding and utilizing deep neural networks trained with noisy labels. In: International Conference on Machine Learning. PMLR; 2019. p. 1062–1070.

[35] Akritidis L, Fevgas A, Manolopoulos Y. An unsupervised distance-based model for weighted rank aggregation with list pruning. Expert Syst Appl 2022;202:117435. [CrossRef]

[36] Guo L, Fang W, Zhao Q, Wang X. The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality. Comput Ind Eng 2021;161:107598. [CrossRef]

[37] Dutta P, Paul S, Cengiz K, Anand R, Kumar A. A predictive method for emotional sentiment analysis by deep learning from EEG of brainwave dataset. In: Artificial intelligence for neurological disorders. Amsterdam: Elsevier; 2023. p. 25–48. [CrossRef]

[38] Dutta P, Paul S, Anand R, Majumder M. A predictive method for emotional sentiment analysis by machine learning from electroencephalography of brainwave data. In: Implementation of smart healthcare systems using AI, IoT, and blockchain. Amsterdam: Elsevier; 2023. p. 109–130. [CrossRef]

[39] Dutta P, Pal S, Kumar A, Cengiz K. Artificial intelligence for cognitive modeling: theory and practice. Boca Raton (FL): Chapman and Hall/CRC; 2023. [CrossRef]

[40] Dutta P, Paul S, Kumar A. Comparative analysis of various supervised machine learning techniques for the diagnosis of COVID-19. In: Electronic devices, circuits, and systems for biomedical applications. Amsterdam: Elsevier; 2021. p.521–540. [CrossRef]

[41] Dutta P, Paul S, Obaid AJ, Mukhopadhyay K. Feature selection based artificial intelligence techniques for the prediction of COVID-like diseases. J Phys Conf Ser 2021;1963:012167. [CrossRef]

[42] Dutta P, Paul S, Sadhu A, Jana GG, Bhattacharjee P. Performance of automated machine learning based neural network estimators for the classification of PCOS. In: Bhattacharyya S, Banerjee JS, De D, Mahmud M, editors. Intelligent human-centered computing. Singapore: Springer Nature; 2023. p. 65–73. [CrossRef]

[43] Karadurmuş E, Göz N, Taşkın N, Yüceer M. Bromate removal prediction in drinking water by using the least squares support vector machine (LS-SVM). Sigma J Eng Nat Sci 2020;38:2145–2153.

[44] Alp S, Yiğit OE, Öz E. Prediction of bist price indices: a comparative study between traditional and deep learning methods. Sigma J Eng Nat Sci 2020;38:1693–1704.

[45] Güneş U, Başhan V, Karakurt AS. Predicting tanker main engine power using regression analysis and artificial neural networks. Sigma J Eng Nat Sci 2023;41:216–225. [CrossRef]

[46] Das O. Prediction of the natural frequencies of various beams using regression machine learning models. Sigma J Eng Nat Sci 2023;41. [CrossRef]

[47] Gupta P, Kumar S, Singh YB, Singh P, Sharm SK, Rathore NK. The impact of artificial intelligence on renewable energy systems. NeuroQuantology 2022;20:5012.

[48] Kumar R, Thakur MA, Rathore NK. Optimizing smart manufacturing with IoT integration and leveraging machine learning analysis. NeuroQuantology 2022;20:1620.

[49] Dutta P, Paul S, Shaw N, Sen S, Majumder M. Heart disease prediction: a comparative study based on a machine-learning approach. In: Artificial intelligence and cybersecurity. Boca Raton (FL): CRC Press; 2022. p. 1–18. [CrossRef]

[50] Dutta P, Paul S, Jana GG, Sadhu A. Hybrid genetic algorithm random forest algorithm (HGARF) for improving the missing value imputation in hepatitis medical dataset. In: 2023 International Symposium on Devices, Circuits and Systems (ISDCS). Piscataway (NJ): IEEE; 2023. p. 1–5. [CrossRef]