Alzheimer's Disease Prediction

Ibtihal Alfayez
Computer Science
King Saud University
Riyadh, Saudi Arabia
443200946@student.ksu.edu.sa

Haifa Almakhdoub
Computer Science
King Saud University
Riyadh, Saudi Arabia
443200820@student.ksu.edu.sa

Leen Almajed
Computer Science
King Saud University
Riyadh, Saudi Arabia
443200636@student.ksu.edu.sa

Ruyuf Abu Qarnayn
Computer Science
King Saud University
Riyadh, Saudi Arabia
443201123@student.ksu.edu.sa

Layan Alamri
Computer Science
King Saud University
Riyadh, Saudi Arabia
443200723@student.ksu.edu.sa

Abstract—**Alzheimer’s disease (AD) presents a major challenge in healthcare, requiring effective and reliable diagnostic tools. This research aims to utilize data mining techniques to analyze demographic, clinical and cognitive data to improve AD classification. Models such as Decision Trees, Naïve Bayes, and Neural Networks were implemented and evaluated using precision, recall, accuracy and F1-score. Among these, Decision Trees showed the most consistent and balanced performance across all metrics effectively recognizing patterns associated with AD.**

Keywords—Alzheimer’s disease, data mining, Decision Trees, Naïve Bayes, Neural Networks, classification metrics.

I. INTRODUCTION

Alzheimer’s disease (AD) is a condition that slowly damages the brain, leading to memory loss, confusion, and changes in behavior [1]. Early detection is critical for managing the disease but traditional methods like cognitive tests and clinical evaluations often rely on subjective judgements. This lack of accuracy creates a challenge in identifying AD in its early stages where intervention is most crucial.

Recent advancements in data mining have introduced new possibilities for analyzing complex data to identify patterns and predict outcomes, offering a more sophisticated approach that could address the limitations of current diagnostic methods.

In this study, we explore how machine learning could help predict AD. The main goal is to demonstrate how data analysis can support healthcare, especially in diagnosing Alzheimer’s early. By studying the links between different factors, we hope to find insights that could help doctors and researchers build better tools for early detection.

II. LITERATURE REVIEW

Advancements in machine learning (ML) have opened new pathways in medical research, particularly in the early diagnosis of Alzheimer’s disease (AD). The development of predictive models using ML algorithms has provided tools that support diagnostic processes by analyzing complex clinical data. This section reviews key studies in this domain, focusing on the methodologies and models that contribute to improved accuracy in AD prediction.

The study in [2], a range of machine learning algorithms was used to analyze large datasets and identify patterns related to AD, aiming to enhance diagnostic accuracy and efficiency. The study utilized a dataset sourced from the open-access platform Kaggle [3], which comprised 35 distinct features relevant to AD diagnosis. The study began with data preprocessing, applying Spearman correlation for feature selection to refine the dataset from 35 to the 13 most relevant features, which improved computational efficiency. Among the models evaluated, including k Nearest Neighbors, Naive Bayes, Decision Trees, and Ensemble methods, the Ensemble model achieved the highest predictive accuracy at 94.07 percent. By combining the strengths of multiple models, the Ensemble method demonstrated a greater ability to detect complex patterns in the data than individual models.

The researchers in [4] utilized the Naive Bayes classifier (NB), Model-Averaged Naive Bayes (MANB), and Feature-Selected Naive Bayes (FSNB) algorithms to predict Alzheimer's disease. These models used neuroimaging data and binary brain properties from health records to distinguish Alzheimer's patients from healthy individuals. The MANB algorithm, which accounts for dependencies between features, achieved the highest area under the curve (AUC) score of approximately 0.72, indicating its effectiveness in Alzheimer’s prediction.

The study [5] utilized decision trees as one of the machine learning techniques to predict Alzheimer’s disease in its early stages. The authors worked with the OASIS dataset[6] , which includes MRI imaging data and clinical information from patients, focusing on features like age, education, brain volume, and cognitive scores. Using decision trees, the data was split into subsets based on feature values, creating a model to classify patients as either demented or non-demented. Alongside decision trees, other models like SVM, Random Forests, XGBoost, and Voting classifiers were also implemented to compare classification effectiveness in Alzheimer’s prediction.

In [7] it explores machine learning algorithms to improve the diagnosis and early prediction of Alzheimer’s disease (AD). This paper focused more on the model explainability and used a large dataset that consists of 169,408 records and 1024 features obtained from the National Alzheimer's Coordinating Centre[8].The algorithms used in this paper were: Applied Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and k-Nearest Neighbors (KNN). Notably, support vector machine (SVM) models exhibited high performance when tested on an external dataset. SVM achieved a high F1 score of 98.9% for binary classification and 90.7% for multiclass classification. Furthermore, SVM was able to predict AD progression over a 4-year period, with F1 scores reaching 88% for binary task and 72.8% for multiclass task. To enhance model explainability, two rule extraction approaches were applied: class rule mining and stable and interpretable rule set for classification model. to assist domain experts in understanding the key factors involved in AD development.

III. DATASET AND ATTRIBUTES

This dataset contains extensive health information for 2,149 patients, with 35 attributes that cover various aspects such as demographics, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and Alzheimer's Disease diagnosis [3]. It is ideal for researchers and data scientists interested in exploring factors associated with Alzheimer's, developing predictive models, and conducting statistical analyses to gain insights into the disease.

Table 1 Dataset Description

Attribute	Description	Type
PatientID	A unique identifier for each patient	Nominal
Age	Age of the patient	Ratio
Gender	Gender of the patient (0 represents male,1 represents female)	Binary
Ethnicity	Ethnicity of the patient, coded as follows: 0: Caucasian 1: African American 2: Asian 3: Other	Nominal
EducationLevel	Education level of the patient, coded as follows: 0: None 1: High School 2: Bachelor's 3: Higher	Ordinal
BMI	Body Mass Index of the patient	Ratio
Smoking	Smoking status (0 indicates no, 1 indicates yes)	Binary
AlcoholConsumption	Weekly alcohol consumption in units, ranging from 0 to 20.	Ratio
PhysicalActivity	Weekly physical activity in hours, ranging from 0 to 10.	Ratio
DietQuality	Diet quality score, ranging from 0 to 10.	Interval
SleepQuality	Sleep quality score, ranging from 4 to 10.	Interval
FamilyHistoryAlzheimers	Family history of Alzheimer's disease (0 for no and 1 for yes)	Binary
CardiovascularDisease	Presence of cardiovascular disease (0 for no and 1 for yes)	Binary
Diabetes	Presence of diabetes (0 for no and 1 for yes)	Binary
Depression	Presence of depression, (0 for no and 1 for yes)	Binary
HeadInjury	History of head injury (0 for no and 1 for yes)	Binary
Hypertension	Presence of hypertension (0 for no and 1 for yes)	Binary
SystolicBP	Systolic blood pressure, ranging from 90 to 180 mmHg.	Ratio
DiastolicBP	Diastolic blood pressure, ranging from 60 to 120 mmHg.	Ratio
CholesterolTotal	Total cholesterol levels, ranging from 150 to 300 mg/dL.	Ratio
CholesterolLDL	Low-density lipoprotein cholesterol levels, ranging from 50 to 200 mg/dL.	Ratio
CholesterolHDL	High-density lipoprotein cholesterol levels, ranging from 20 to 100 mg/dL	Ratio
CholesterolTriglycerides	Triglycerides levels, ranging from 50 to 400 mg/dL.	Ratio
MMSE	Mini-Mental State Examination score, ranging from 0 to 30, with lower scores indicating cognitive impairment.	Interval
FunctionalAssessment	Functional assessment score, ranging from 0 to 10, with lower scores indicating greater impairment.	Interval
MemoryComplaints	Presence of memory complaints (0 for no and 1 for yes)	Binary
BehavioralProblems	Presence of behavioral problems (0 for no and 1 for yes)	Binary
ADL	Activities of Daily Living score, ranging from 0 to 10, with lower scores indicating greater impairment.	Interval
Confusion	Presence of confusion (0 for no and 1 for yes)	Binary
Disorientation	Presence of disorientation (0 for no and 1 for yes)	Binary
PersonalityChanges	Presence of personality changes (0 for no and 1 for yes)	Binary

DifficultyCompletingTasks	Presence of difficulty completing task (0 for no and 1 for yes)	Binary
Forgetfulness	Presence of forgetfulness (0 for no and 1 for yes)	Binary
Diagnosis	Diagnosis status for Alzheimer's Disease	Binary
DoctorInCharge	Confidential information, about the doctor in charge, with "XXXConfid" as the value for all patients.	Nominal

IV. Data Mining Techniques and Algorithms

The selected data mining approach is classification, specifically Decision Tree, Naive Bayes and Multilayer Feed-forward Neural Network all of which are supervised learning methods.

A. Decision Tree

Decision Trees are widely used for supervised learning tasks such as classification. The CART (Classification and Regression Trees) algorithm constructs these trees by evaluating splits at each node to optimize a criterion. For classification problems, one such criterion is cross-entropy, which measures the impurity of a node based on the class distribution.

The entropy at a node is calculated using the following equation.

$$\mathrm{E} = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

where p_i is the proportion of samples belonging to class i and k is the number of classes.

Entropy helps determine the uniformity of the data at a node:

High entropy: Classes are evenly distributed (high impurity).

Low entropy: One class dominates (low impurity).

CART uses information gain to decide the best split for a node. Information gain measures the reduction in entropy after splitting a node. It is defined as:

$$\text{Information Gain} = \mathrm{E(Parent)} - \sum_{i=1}^{k} \frac{n_i}{N} \cdot E(Child_i)$$

The algorithm evaluates all possible splits and selects the one with the highest information gain.

B. Naïve Bayes

The Naive Bayes classifier is a probabilistic model that uses Bayes' Theorem to predict the likelihood of a class given a set of features[9]. Specifically, Bayes' Theorem provides a mathematical foundation for calculating the posterior probability P(C|X) of a class C given an observation X (a set of features). The equation is defined below.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where P(C|X) denotes the posterior probability of class C given X, P(X|C) is the likelihood of observing X given C, P(C) is the prior probability of C, and P(X) is the probability of observing X.

An important aspect in the Naïve Bayes model is the "naive" assumption of conditional independence. This assumption allows the overall likelihood P(X|C) to be decomposed as the product of individual probabilities for each feature as calculated in the following equation.

$$P(X|C) = \prod_{i=1}^{n} P(X_i|C)$$

This simplification reduces computational complexity of the model and allows each feature's likelihood to be calculated independently.

Among the different variants of Naive Bayes classifiers, Gaussian Naive Bayes is commonly used for continuous data as it assumes each feature is normally distributed within each class. For a continuous feature X, the conditional probability $P(X_i|C)$ is modeled by the Gaussian (or normal) distribution, defined as:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

where μ and σ are the mean and standard deviation of the feature Xi for class C. By modeling each continuous feature with a Gaussian distribution, the classifier can compute the likelihood of observing specific feature values for each class, which supports classification even in cases where feature values vary widely.

C. Neural Network

A neural network is a computational model inspired by the human brain, designed to learn patterns and labels from data. A neural network consists of layers of interconnected neurons. Each neuron sums a weight to its inputs, adds a bias, and applies an activation function to produce an output, The output of a single neuron is expressed as follow:

$$f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

where Xi are the inputs, Wi are the weights, b is the bias term, and f is the activation function, for a neural network layer, this generalizes to:

$$f\left(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}\right)$$

Neural networks are widely used in data mining, enabling tasks such as classification, clustering, and prediction, and their ability to adapt to diverse data sets [10].

V. Correlation

The relationship between the different attributes was analyzed to determine the associations. Correlation analysis was performed to measure the strength and direction of the linear relationships between the attributes. The results below reveal the highest positive correlation and the highest negative correlation.

Highest Positive Correlation:

The strongest positive relationship observed between "Memory Complaints" and "Diagnosis", with a correlation value of 0.3067, which indicates a moderate positive correlation. This indicates that individuals reporting higher memory complaints are more likely to receive a diagnosis, emphasizing the potential diagnostic relevance of subjective memory assessments.

Highest Negative Correlation:

The strongest negative relationship identified between “Functional Assessment” and “Diagnosis”, with a correlation value of -0.3649, which indicates a moderate inverse correlation. This finding implies that as functional assessment scores increase, the likelihood of receiving certain diagnoses decreases. Figure 1 shows the correlation between each attribute.

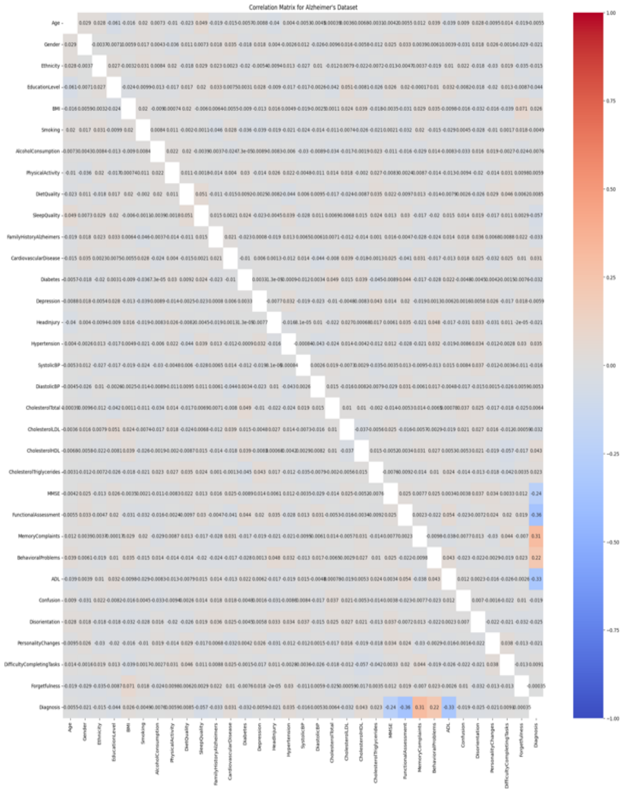

Figure 1 Correlation Matrix for Alzheimer's Dataset

VI. Preprocessing

The quality and reliability of the dataset significantly influence the outcomes of mining operations [11]. To make sure the dataset was thoroughly prepared for the intended tasks, the following preprocessing steps were performed. The dataset was reviewed to ensure there were no missing values and this process confirmed that all entries were complete. It was also examined for duplicate rows, with none being identified. As part of this step, attributes like "PatientID" and "DoctorInCharge" were removed, as they were not relevant to the machine learning algorithms used in this study. For data transformation, categorical attributes in the dataset were encoded into a numerical format using one-hot encoding. As for numerical attributes in the dataset, they were standardized to a uniform scale. In addition, attributes with natural or predefined limits which are also known as bounded features, were scaled using the Min-Max Scaling technique to fit within a range of [0, 1].

The quality and reliability of the dataset significantly influence the outcomes of mining operations [11]. To make sure the dataset was thoroughly prepared for the intended tasks, the following preprocessing steps were performed. The dataset was reviewed to ensure there were no missing values and this process confirmed that all entries were complete. It was also examined for duplicate rows, with none being identified. As part of this step, attributes like "PatientID" and "DoctorInCharge" were removed, as they were not relevant to the machine learning algorithms used in this study. For data transformation, categorical attributes in the dataset were encoded into a numerical format using one-hot encoding. As for numerical attributes in the dataset, they were standardized to a uniform scale. In addition, attributes with natural or predefined limits which are also known as bounded features, were scaled using the Min-Max Scaling technique to fit within a range of [0, 1].

VII. Model Testing and Evaluation

A. Decision Trees

The implementation of the decision tree was performed using the DecisionTreeClassifier function from Scikit-learn's tree module, with the criterion set to entropy. The criterion determines how the decision tree evaluates splits. In our case, using entropy resulted in better performance metrics compared to the Gini index, which measures impurity based on squared probabilities of class labels. The model achieved an overall accuracy of 91%, with a precision, recall, and F1-score of 0.93 for Class 0 and 0.88 for Class 1. The macro and weighted averages for precision, recall, and F1-score were all 0.91, indicating balanced performance across both classes. Figure 2 shows the confusion matrix for the decision tree model.

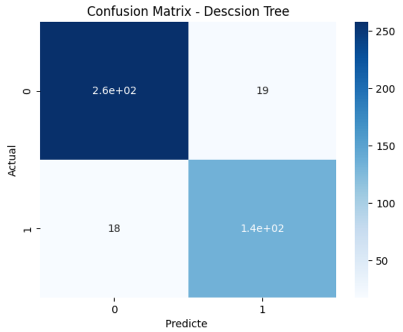

Figure 2 Decision Tree Confusion Matrix

B. Naïve Bayes

The Naïve Bayes model used in out implementation is Gaussian Naive Bayes (GaussianNB), which is a part of Scikit-learn library. The model was trained using the fit method on the training set then evaluated using the method predict on the test set.

The Naïve Bayes model achieved an overall accuracy of 83% with a weighted average precision, recall, and F1-score of 83%, Figure 3 shows the confusion matrix for the Naïve Bayes model.

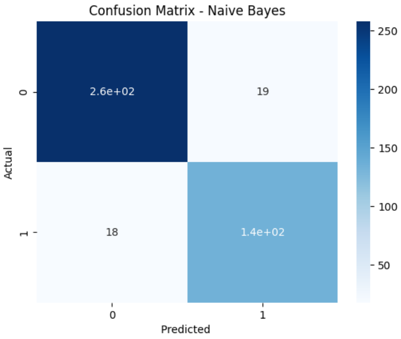

Figure 3 Naïve Bayes Confusion Matrix

C. Neural Network

The neural network model is implemented using Scikit-learn's MLPClassifier, which provides a flexible framework for building feedforward neural networks suitable for classification tasks. It consists of three hidden layers, each with 10 neurons. The model utilizes the logistic activation (sigmoid) function, which outputs probabilities between 0 and 1, making it ideal for binary class classification, the learining rate was set to 0.001 by default. The optimizer is Adam, a variant of stochastic gradient descent that adapts learning rates during training, paired with the cross-entropy loss function for optimizing classification performance. The model is configured to train for a maximum of 10,000 iterations, ensuring sufficient time for convergence.

The model achieved an overall accuracy of 83%, demonstrating reliable performance in predicting the target variable. The weighted averages for precision, recall, and F1-score were all 0.83. Figure 4 shows the confusion matrix for the Neural Network.

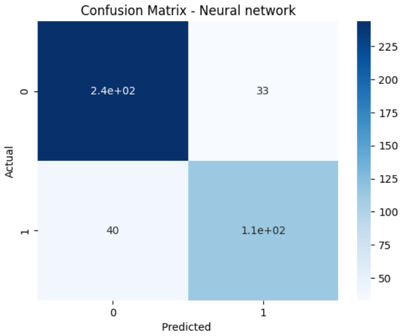

Figure 4 Neural Network Confusion Matrix.

VIII. Comparison and Discussion

In our work we used three classification algorithms: decision trees, naïve bayes and neural networks.

Table 2 Comparison of the three classification models used in this research

Model	Precision	Recall	F1-Score	Accuracy
Decision Trees	0.91	0.91	0.91	0.91
Naïve Bayes	0.83	0.83	0.83	0.83
Neural Network	0.83	0.83	0.83	0.83

From table 2 we can see that the best classification model with the highest accuracy is the decision tree model with an accuracy of 91%. for the naïve bayes and neural network classification models they have the same accuracy of 83%. It is worth noting that the number of hidden layers, number of max iterations and the activation function used here can play a huge role in the performance of the neural network classifier.

IX. Conclusion

This study examined how machine learning models—Decision Trees, Naïve Bayes, and Neural Networks—can help predict Alzheimer’s Disease (AD). Among the models tested, Decision Trees performed the best with an accuracy of 91%, providing reliable results across all key metrics. Naïve Bayes and Neural Networks both achieved an accuracy of 83%, showing potential for further improvement.

Our analysis also revealed that certain features, like memory complaints and functional assessment scores, are strongly linked to AD diagnosis. This highlights how machine learning can support doctors by making early detection more accurate and effective.

X. References

[1] A. P. Porsteinsson, R. S. Isaacson, S. Knox, M. N. Sabbagh, and I. Rubino, “Diagnosis of Early Alzheimer’s Disease: Clinical Practice in 2021,” *J Prev Alzheimers Dis*, pp. 1–16, 2021, doi: 10.14283/jpad.2021.23.

[2] A. Samad and E. Samet Aydı, “Rapid Alzheimer’s Disease Diagnosis Using Advanced Artificial Intelligence Algorithms,” *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1760–1768, Jul. 2024, doi: 10.38124/ijisrt/IJISRT24JUN1915.

[3] R. El Kharoua, “Alzheimer’s Disease Dataset.” Accessed: Nov. 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

[4] W. Wei, S. Visweswaran, and G. F. Cooper, “The application of naive Bayes model averaging to predict Alzheimer’s disease from genome-wide data,” *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 370–

375, Jul. 2011, doi: 10.1136/amiajnl-2011-000101.

[5] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, "Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models," *Front Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.853294.

[6] " Open Access Series of Imaging Studies (OASIS) ." Accessed: Nov. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers

[7] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for Alzheimer's disease classification," *Sci Rep*, vol. 14, no. 1, p. 2637, 2024, doi: 10.1038/s41598-024-51985-w.

[8] "NACC dataset." Accessed: Nov. 18, 2024. [Online]. Available: https://naccdata.org/

[9] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft comput*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[10] P. Gaur, "Neural networks in data mining," *International Journal of Electronics and Computer Science Engineering*, 2012.

[11] F. Berzal and N. Matín, "Data mining: concepts and techniques ," *ACM SIGMOD Record*, vol. 31, no. 2, pp. 66–68, Jun. 2002, doi: 10.1145/565117.565130.