Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree

Hala Ahmed, Hassan Soliman, Mohammed Elmogy [*]

Information Technology Dept., Faculty of Computers and Information, Mansoura University, Mansoura, P.O.35516, Egypt

ARTICLE INFO

Keywords:
AD
Alzheimer's disease
Classification
Prediction
Diagnosis
SNPs
Single nucleotide polymorphisms
GBT
Gradient boosting tree
Boruta feature selection
Information gain feature selection

ABSTRACT

Alzheimer's disease (AD) is a degenerative disorder that attacks nerve cells in the brain. AD leads to memory loss and cognitive & intellectual impairments that can influence social activities and decision-making. The most common type of human genetic variation is single nucleotide polymorphisms (SNPs). SNPs are beneficial markers of complex gene-disease. Many common and serious diseases, such as AD, have associated SNPs. Detection of SNP biomarkers linked with AD could help in the early prediction and diagnosis of this disease. The main objective of this paper is to predict and diagnose AD based on SNPs biomarkers with high classification accuracy in the early stages. One of the most concerning problems is the high number of features. Thus, the paper proposes a comprehensive framework for early AD detection and detecting the most significant genes based on SNPs analysis. Usage of machine learning (ML) techniques to identify new biomarkers of AD is also suggested. In the proposed system, two feature selection techniques are separately checked: the information gain filter and Boruta wrapper. The two feature selection techniques were used to select the most significant genes related to AD in this system. Filter methods measure the relevance of features by their correlation with dependent variables, while wrapper methods measure the usefulness of a subset of features by training a model on it. Gradient boosting tree (GBT) has been applied on all AD genetic data of neuroimaging initiative phase 1 (ADNI-1) and Whole-Genome Sequencing (WGS) datasets by using two feature selection techniques. In the whole-genome approach ADNI-1, results revealed that the GBT learning algorithm scored an overall accuracy of 99.06% in the case of using Boruta feature selection. Using information gain feature selection, the proposed system achieved an average accuracy of 94.87%. The results show that the proposed system is preferable for the early detection of AD. Also, the results revealed that the Boruta wrapper feature selection is superior to the information gain filter technique.

1. Introduction

Alzheimer's disease (AD) is a sort of dementia and is regarded as the most common type [1]. AD is a brain disorder that leads to memory loss, cognitive and intellectual impairments, and the ability to influence social activities and decision-making. One of the critical studies is identifying complex diseases related to genetic variants associated with the human genome. Genome Wide Association Studies (GW-AS) aim to identify genetic variants, especially complex diseases related to single nucleotide polymorphisms (SNPs). SNPs are the most common type of genetic variation among people. The disease occurs when one of the nucleotides of Adenine (A), Thymine (T), Cytosine (C), or Guanine (G) differ in their DNA sequence. The SNPs are the primary goal of genetic association studies to determine the most associated SNPs with common and complex diseases [2] (see Table 1).

AD shows several features for clinical and pathological symptoms that lead to disease division into three main stages: early (mild), intermediate (moderate), and late (severe). It is greatly beneficial to diagnose the disease in its early stages due to several factors: (a) maximizing the benefit of innovating new treatment strategies aimed at changing the impact of the disease in its early stage, (b) maintaining patients' daily functions as much as possible by slowing the effect of worsening disease symptoms, and (c) providing long-term care and medical costs for both patients and governments. Nevertheless, diagnosing the disease in its early stages is regarded as a hard challenge in this field of research due to several reasons involving the late appearance of pathological features related to the disease. These features generate in the body for 10–15 years before becoming visible. It means that the clinical diagnosis of the disease is performed 10–15 years post contraction of the disease [3].

Symptoms appear at the age of 65, and the spread of disease with age

[*] Corresponding author.
E-mail address: melmogy@mans.edu.eg (M. Elmogy).

https://doi.org/10.1016/j.compbiomed.2022.105622
Received 10 May 2021; Received in revised form 25 March 2022; Accepted 29 March 2022
Available online 24 May 2022

Table 1
An overview of complex brain diseases using different ML methods.

Authors	Disease	Methods	Results	Problem
Michael et al. [18]	AD	DM methods	The results described that RF and MDR are powerful methods than existing methods for detecting genetic interactions.	Their work need adding other modalities may improve the prediction accuracy.
Abd El Hamid et al. [19]	AD	SVM	Results shown that RBF kernel is used with SVM trained model has a better linked with AD and perform good accuracy of 76.70%.	Choose different genes based on selection methods are needed to investgate more higher genes that may aid in discover new biomarkers of disease using other ML algorithms.
Spencer et al. [21]	ASD	FPM algorithms & contrast mining	Including 193 novel autism candidates as significant associations from connected 286 genes.	It is a challenge for FPM to store many combinations of items as a memory requirement problem.
Boutorh et al. [22]	breast cancer	hybrid intelligent technique based on (ARM) and NN	Their model has achived an accuracy up to 90%.	The classification performance need more enhancements.
Narayanan et al. [23]	lung cancer	SVM	CAD perform Sensitivity with 82.82%.	They should optimize the feature set for SVM classification
Hu et al. [24]	AD	SMR	The accuracy of the results is confirmed.	The GWAS still has limitations. The strategy is based on the "common disease" hypothesis, which misses rare variants which can play a more important role in causing diseases.
Mukherjee et al. [25]	AD	ML algorithm	They explained that their ranked genes show significant enrichment for AD.	Their studies need to focus on specific genes and pathways that are driving disease etiology

increases sharply. Besides, AD is considered the most common kind of dementia in the onset of genetic disease factors. Although it is not the primary cause of the disease, a specific gene can play an important role. However, the symptoms can increase due to these factors [4,5]. Other factors affecting the disease are smoking and alcohol. Complete loss of memory, impairments of movements, misplacing things, verbal communication difficulties, and abnormal mood swings are defining symptoms of the disease. If it is not diagnosed initially, the disease's severity increases, as shown in Fig. 1 [6].

Consequently, diagnosis and treatment of such disease could be made in early stages and with higher accuracy for detection. So far, the most critical risk factor is Apolipoprotein E (APOE) gene that is confirmed or listed in the AlzGene database. For genetic studies, machine learning (ML) techniques have been applied to explore the genetic variants that have the most association with complex diseases. The ML classifier was utilized to classify patients into AD, mild cognitive impairment (MCI), and control subjects to discover new genetic biomarkers for AD progression.

Despite the success of standard artificial intelligence (AI) techniques for gene expression data analyses, it has become apparent that it is challenging to analyze large-scale data using the only one-standard smart approach. If a conventional classifier is applied to gene expression for disease diagnosis to classify a sample based on all the variables, low accuracy would be expected. Microarray data is a well-known phenomenon that produces many features and a relatively small number of samples known in ML as the curse of the dimensionality problem. A small number of selection techniques of the informational features became essential to reduce computational costs, aid in identifying a small subset of genes that are biologically relevant to different diseases, and obtain the required prediction accuracy [7]. In general, defining a feature aims to remove inconvenient and duplicate features, thus making the classifier work better as a diagnostic model. With the increasing availability of more and different types of omics data, ML methods have become more frequent. One of the challenges for ML approaches is predicting genomic features, especially regulatory regions prediction, as they are difficult to predict by applying contemporary methods. Accordingly, ML has been applied to predict the sequence properties of proteins associated with DNA and RNA, enhancers, and other regulatory regions [8–10].

SNPs work as pointers in the association and linkage studies to identify the genome's part in a particular disease [11,12]. Polymorphs found in the same coding and organizing regions may be another contributor to diseases. A non-synonymous SNP holds great interest for researchers because they cause amino acid substitution. A vast number of variations in amino acids lead to genetic diseases. SNPs may have critical biological effects that have been shown in several studies, such as being related to complex diseases. SNP is a sequence variation of DNA from a single nucleotide change in the genome, and it is considered the most common genetic variation [13]. So, this paper presented a comprehensive framework for early diagnosis of AD based on SNPs analysis to improve the accuracy and detect the most significant genes or SNPs. Two methods were used for feature selection: Boruta and information gain (IG). The gradient boosting tree (GBT) was used to classify normal, MCI, and AD. Besides, it detected most SNPs associated with the disease. In order to implement these ML algorithms, the appropriate preprocessing steps must be determined in advance. Classification studies that use ML generally need more essential steps: feature extraction, feature selection, dimensionality reduction, and feature-based classification algorithm selection. Specialized knowledge is required for these procedures and multiple stages of improvement. Reproduction of these methods is an important issue.

The main contributions of our work can summarize in the following points:

- The data were preprocessed to enhance their quality, which increased the performance of the diagnosis process. First, the outliers were handled by substituting their values with the median. Second, the missing data were handled by using the median. Third, the attributes were normalized to provide a unique scale for all used features. Finally, phenotypic information processing is defined for NC, MCI, and AD.
- Two feature selection techniques were used: Boruta and IG feature selection, to get the most significant AD features provided by the experimental results. The results revealed that the Boruta feature selection achieved the highest performance in diagnoses.
- Our work's main concern or novelty is detecting the disease in its early stages by identifying the most significant genes and SNPs that cause the disease. Seven candidate genes were found, which are CLU, ABCA 7, APOE, BINI, CRI, CD2AP, and CD33. They are the most highly associated genes with AD, identified as the most significant

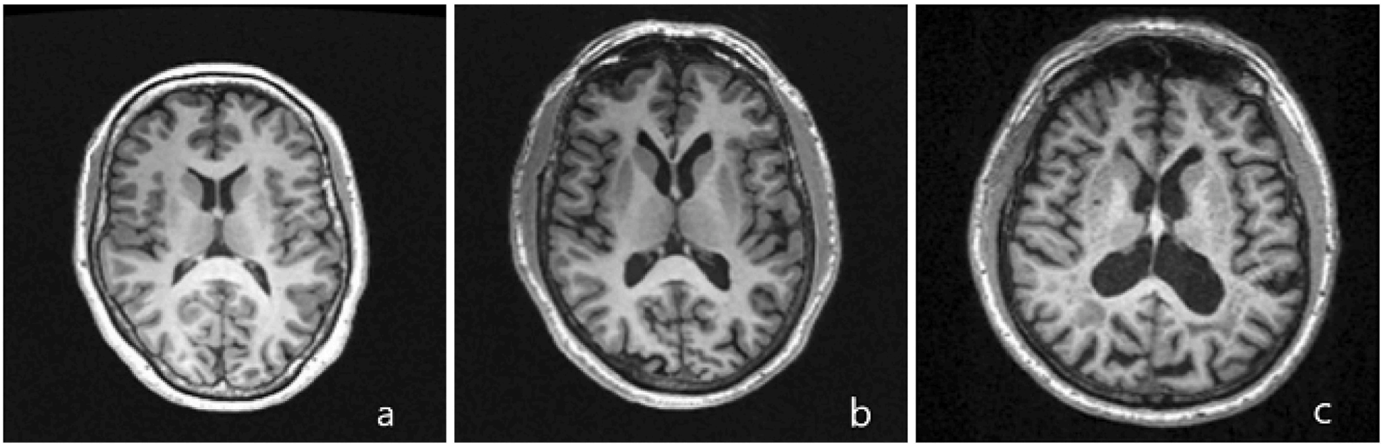

Fig. 1. The T1-weighted MRI imaging for different stages of AD patient (a) The healthy brain, (b) MCI brain, and (c) AD brain.

SNPs. These SNPs are regarded as critical biomarkers for the disease. A number of common SNPs are obtained as significant SNPs for early detection of AD as follows: rs512941, rs2074451, rs429358, rs1595816, rs17014396, rs9296559, rs2007-332, and rs204I992.

- A comparison was provided between our framework and other state-of-the-art techniques. Our framework has the highest accuracy compared with other techniques.
- A well-tuned GBT classifier was used to train and test our model in the fifth stage of the proposed framework system by hyper-parameterizing their parameters.

The remainder of this paper is divided into five sections. Section 2 introduces the related work, current limitations of recent research, and how our proposed framework overcame these limitations. Section 3 provides the framework of the proposed system in more detail for each used technique, briefly explaining all of them. Section 4 discusses the results obtained from the proposed system. Section 5 presents a discussion on the experimental results. Finally, the conclusion and suggestions for future work are presented in Section 6.

2. Related work

There is an urgent need to diagnose brain diseases by identifying genetic biomarkers to provide accurate detection. So, this section introduces a comprehensive review of genome sequencing analysis for discovering complex genes related to genetic brain diseases [14–17]. Most of the genetic variations in human genomes are contributed by SNPs. Many complex and common diseases are related to SNPs like AD. Early diagnosis can be improved by identifying SNP biomarkers at different loci for related diseases. Complex diseases investigation of genetic variants in the human genome is considered one of the most crucial study subjects [6,7]. For example, Mikhail et al. [18] aimed to measure the effect of the association at the genome level by studying SNPs in AD. Data mining (DM) methods have been tested. The used data was taken from the ADNI database. Multiple models were applied, such as linear regression (LR), random forest (RF), and multifactor dimensionality reduction (MDR). The results show that using RF and MDR is more effective. The MDR model achieved the overall sensitivity in all comparisons using only 3 SNPs. The LR yielded high specificity in two comparisons (cognitively normal (CN) versus MCI and MCI versus AD). At the same time, MDR provided the best AD specificity compared to NC. Several significant SNPs associated with MCI and AD have been identified [18].

Abd El Hamid et al. [19] presented a method to detect complex diseases related to genetic biomarkers. The goal of their work was to determine the important forms of SNP associated with AD. In their work, a sequential minimal optimization (SMO) model trained using different kernels has been proposed to identify the most important forms of polymorphism associated with AD. Significantly AD-related SNPs are identified in many genes. Methods of feature selection are essential to eliminate some of the unimportant polymorphisms to enhance classification performance. They used two feature selection methods to determine the most significant SNPs. The radial base function kernel (RBF) is used to train SMO on the best-selected polymorphisms with a higher association with AD disease and better accuracy. They concluded that SNPs identified in AD's early stages are essential biomarkers as they help enhance medical diagnostic methods and discover the causes of the disease [20].

Spencer et al. [21] presented a heritable genotype. Their system consists of five stages. The preprocessing is the first step to impute the missing SNPs for testing and select the most significant SNPs. The second stage is the division of the population. With each subgroup procedure of genome-wide, prioritization is used as a primary association. The third stage is genome-wide prioritization by returning to the question: how do they decide which sets of SNPs to test? The answer to this question is the frequent pattern mining (FPM) algorithms utilizing the minimum threshold support. The filter will be performed for SNP sets according to their prevalence among the affected population. The fourth stage is FPM, one of the DM techniques that excel most in feature combinations to identify the most common occurrence repeatedly. The data is required to be interpreted into binary, with the two referring states in a person, which are the presence or absence of the item in FPM, to highlight potential interactions between variants. Finally, the last stage is the UICsup, which is a contrast mining utilization.

Boutorh et al. [22] introduced a hybrid technique that depends on association rule mining (ARM) and neural networks (NNs), which used an evolutionary algorithm (EA) that was presented to handle the problem of dimensionality for breast cancer diagnosis. ARM is performed to show the most critical features and decrease the dimensionality by extracting associations between SNPs, while for efficient classification, NN is used. Their method NN-GEARM had been performed on an SNP dataset for breast cancer. The developed model achieved accuracy up to 90%.

Narayanan et al. [23] presented a study of lung cancer and explored the performance of the support vector machine (SVM) based on a wide range of features. The results showed that the SVM is mathematically more powerful and faster with a wide range of features and is less prone to overtraining than conventional classifiers. Besides, they also offer a computationally efficient approach to selecting SVM features. Results are presented to the publicly available 2016 dataset for lung nodule analysis. Their results showed that the SVM classification method, which used 10-fold validation, was superior to the fisher linear discrimination classifier by 14.8%.

In order to identify locus and genes associated with AD, Hu et al. [24]

introduced a summary mendelian randomization (SMR) approach. They completed ten experiments and collected efficient results from five experiments using two GWAS datasets and five Expression quantitative trait loci (eQTL) datasets. A total of 27 SNPs associated with AD were identified. These SNPs correspond to seven genes. They compared results with known databases to verify the efficiency of their method and the accuracy of results. Three of the seven genes were found to be novel genes, and six of the seven genes are novel genes in the DisGeNET database.

Mukherjee et al. [25] introduced a method to combine multiple data types to extract driven data and analyze results. Then, evidence was aggregated to develop the hypothesis that a gene is the genetic driver of disease. Their work followed two basic stages: (i) From multiple features sets, a general ML framework is presented to discover the few known driver genes key characteristics and to identify the similar feature representations of the driver genes, and (ii) A scheme for flexible ranking with the ability to incorporate external validation into the genome association study summary stats form. Also, they demonstrated the utility of their ML method over two standardized multi-view datasets. Then, they used their method to predict and rank potential drivers of AD.

Nowadays, people face numerous diseases due to their living habits and the conditions of the environment. Therefore, predicting disease at an early stage became an important goal. However, accurate prediction based on symptoms is too difficult for doctors. The correct prediction of disease is the most challenging aim. To overcome this difficulty, using ML would be the best choice. Algorithms in ML are easy to implement and are flexible enough to deal with complex problems with multiple interacting variables.

As described above, some limitations for the current related work can be concluded in the following points. First, the selection of features is the most critical step. The commonly used feature selection is filter methods. Still, it has many disadvantages: they neglect the interaction with the classifier, the features are considered independently for each feature, and neglecting feature dependencies. In addition, it is not clear how to specify the threshold point for rankings to choose only the required features and exclude noise. Second, many studies ignore the preprocessing phase and analysis of data for dealing with missing data. Third, a poor analysis of gene selection or candidate genes. Finally, the main limitation of most models in some ML studies is the overlooking of serious overfitting problems.

So, with the above-mentioned limitations, we developed the proposed framework for the early detection of AD based on SNPs. First, we concentrate on the prepossessing phase to deal with missing values, which is a focus point. Secondly, wrapper feature selection is used to overcome the limitations in filter feature selection to select the essential feature that helps the classifier in classification. Unlike filter methods which use feature-relevant criteria, the wrapper methods are based on the performance of classifiers for obtaining a feature subset. So, the accuracy of the classifier increases in the case of using wrapper feature selection. Finally, overfitting problems are handled by applying different cross-validation techniques on tested data.

3. The proposed framework

Variations in the human genome are an essential factor affecting AD susceptibility. Consequently, discovering genetic biomarkers for complex diseases, including AD, is the goal of the proposed work. The target is to identify the most significant SNPs associated with AD. ML is increasingly being applied in healthcare to build models, develop practice guidelines or refine guidelines for better medical decision making. So, the main objective of the developed framework is to create a comprehensive system for early detection of AD based on SNPs analysis. Our framework consists of five main stages, as shown in Fig. 2. The first stage is dealing with the database. The Alzheimer's Disease Neuroimaging Initiative (ADNI) database is used for this study. The second stage is the preprocessing stage, which includes the data normalization, imputation, and APOE genotyping combination. The third stage is the most crucial stage known as feature selection, which helps enhance the classifier's efficiency. In the fourth stage, GBT classifiers' ML techniques are used. We use a GBT classifier to train and test our model. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. Also, we can merge in the fourth stage, the accurate prediction of three AD cases, and the AD biomarkers are identified in the early stages.

rs113464261, rs769449, rs73504429, APOE112, rs4844-609, rs4732729, rs9331942, rs610932, rs7530069, rs114506-298,

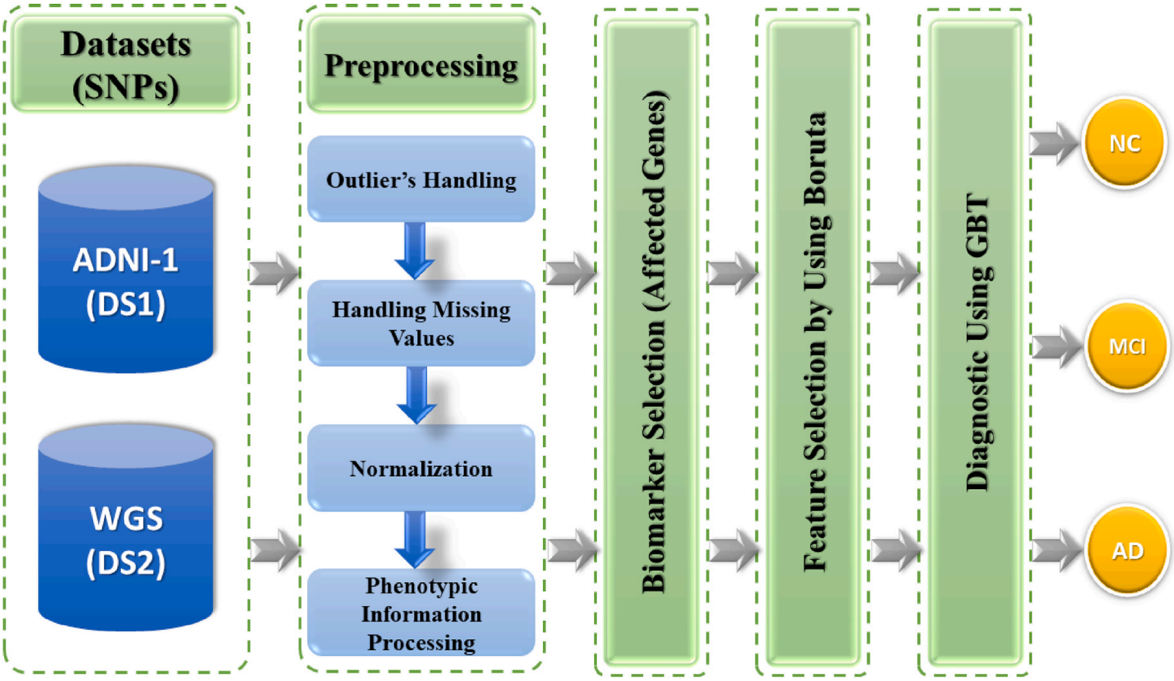

Fig. 2. The proposed framework for detecting and diagnosing of AD in early stages.

rs7929589, and rs611267 were among the SNPs reported to be associated with AD risk in previous studies. Other SNPs were identified to be strongly linked to AD in our work. The SNPs involved rs512941, rs2074451, rs15958-16, rs17014396, rs9296559, and rs2007332. These findings revealed that GBT could be used to identify AD causal SNPs with reasonable accuracy.

3.1. Dataset

Data from the ADNI database were obtained in preparation for this paper. The ADNI study has been divided into several phases, including ADNI-1, ADNI-GO, and ADNI-2. In 2003, the ADNI, led by the principal investigator Michael W. Weiner, was released as a public-private partnership. In this work, there are two used datasets. ADNI phase 1 data is collected from 757 total subjects (214 controls, 366 MCI, and 177 AD cases) [26] the direct link to download ADNI. The Human 610- Quad BeadChip was used for genotyping the ADNI-1 samples resulting in 620, 901 SNPs. WSG includes 812 individuals (321 controls, 442 MCI, and 49 AD cases). WGS samples were genotyped using Illumina Omni 2.5 M, resulting in 2,379,855 SNPs considered big data, including domain, tested with different phenotypes.

3.2. Preprocessing

The data preprocessing stage is an essential step for getting meaningful results. In the proposed framework, the data preprocessing stage consists of four steps: outliers handling, missing values handling, normalization, and phenotypic information processing. First, median imputation is a simple procedure in which the missing entries of the data matrix are estimated using the median of the non-missing values of the particular case or variable (row average or column average), respectively. Missing values were filled with the median of the observed values per variable. As we know, median imputation is suitable when data contains outliers. The outliers are data points lying far away from the majority of other data points. Outliers in the data that are not normally distributed do not require identification. As most statistical tests assume that data are normally distributed, outlier identification should precede data analysis. An outlier is defined as any point of data that lies below 1.5 IQRs of the first quartile (Q1) or above the third quartile (Q3) in a dataset. Also, IQR is the difference between Q3 and Q1, as shown in Eqs. (1) and (2).

$$High = (Q3) + 1.5(IQR) \tag{1}$$

$$Low = (Q1) - 1.5(IQR) \tag{2}$$

Second, the missing data problem can be handled in three ways, but we used the second and third methods to handle data, which gave us the best results.

1. All samples with a missing record are removed before any analysis occurs. This is a reasonable approach when the percentage of removed samples is low so that a possible bias in the study can be discarded. On the other hand, the missing values can be estimated from the incomplete measured data. This approach is known as imputation and is recommended when the adopted data analysis techniques are not designed to work with missing entries. About 80% of the ADNI patients have missing records. Despite this, such patients are discarded in the vast majority of ADNI studies. We know that discarding 80% of the patient's information is a serious concern. So, in our paper, we avoided neglecting missing records by using imputation methods [27,28].
2. Mean imputation is a simple procedure in which the missing entries of the data matrix are estimated using the average of the non-missing values of the particular case or variable (row average or column average) respectively. Missing values are filled with the mean of the observed values per variable. So, genes with zero expression are replaced with imputation methods for the records containing zero expression to keep valuable information. Gene without expression value should be handled across all samples, and this step is considered one of the simplest used preprocessing methods.
3. Median imputation is a procedure in which the data matrix's missing entries are estimated using the median of the non-missing values of the particular case or variable (row average or column average). Missing values are filled with the median of the observed values per variable [29,30]. The median performs well in the case of outliers than mean imputation.

Third, data normalization is performed by changing the range or scale of the data to a range from 0 to 1. The function of data normalization is described by the Min-Max normalization method in Eq. (3), where y' is the value of the feature in the domain of normalized data. In contrast, the original value of the data is y before the operation of normalization is performed. $y_{\max}$ and $y_{\min}$ refer to the largest and the smallest values of all attributes in the data to be normalized, respectively.

$$y^{'} = \frac{y - y_{min}}{y_{max} - y_{min}} \tag{3}$$

Finally, APOE genotyping was combined with the dataset for quantifying the rs7412 and rs429358 SNPs alleles. Also, diagnostic information is included to specify phenotypic information for each subject in the dataset as healthy control, MCI, and AD. APOE and status are specified by two SNPs: rs429358 and rs7412. ADNI-l data contains separate APOE genotyping.

In the ADNI dataset, genetic evaluation of genotyping of APOE is not included. At the time of individual registration, APOE SNPs (rs429358 and rs7412) are genotyped. These two genetic variants define three alleles known as ε_2, ε_3 and ε_4 variants. All participants in the ADNI database obtained these variants. Two SNP genotypes (rs429358 and rs7412) are obtained from three alleles genotype. The dataset is complemented with the APOE genotyping by estimating rs7412 and rs429358 SNPs alleles.

In our paper, we used two versions of the ADNI datasets. The first dataset is the WGS ADNI dataset, which includes only CN and AD. On the other hand, the second dataset is the ADNI-1 which includes three different classes which are CN, MCI, and AD. MCI is an early stage of memory loss or other cognitive ability loss, which is considered an early sign of AD. Therefore, we used the second dataset to make our diagnosis system capable of handling the early signs of the disease. Using the first ADNI dataset, 620 and 903 SNPs were obtained after adding APOE genotyping. However, the number of SNPs was 2, 379, and 857 after adding APOE genotyping in the second dataset. The phenotype representation (0, 1, and 2) is used for CN, MCI, and AD.

3.3. Feature selection

Selecting a subset of related features is known as feature selection (FS), which helps build the model. Basically, there are three methods. First, the wrapper methods perform all possible subsets of the dataset. Then, a classification algorithm is used to induce features of classifiers in each subset. The evaluator uses a search technique, such as random search and depth search, to obtain a subset. Second, the filter method uses an evaluator and ranker to rank all features in predefined data set and arranges attributes. Finally, we delete the lower-ranked feature one by one, so the dominant features can be identified [16]. Third, the feature selection process in the embedded methods is an integral part of the classification model [17].

In this study, we tried to reduce feature dimensions and select significant features that enhance the performance. So, the main goal of using FS is to use only a selected subset of features, which enhances rating performance by deleting unimportant features. To identify the best subset of features among many features by using the ideal technique is known as FS. Finding the accurate subset of features is a critical goal in itself. It is time-consuming to use all datasets of the disease or all features subset in the classification process. In addition, some genes may lead to AD while the remainder genes do not. Hence, knowing which genes have a stronger influence on either or both diseases helps obtain higher accuracy for classification.

3.3.1. *Boruta algorithm*

The Boruta FS algorithm is an RF-based strategy that deletes features that have proven less useful than random investigations. The most used classification algorithm is the RF because it has less calculation time and a free parameter for manual tuning. RF is based on multiple decision trees which gather on weak classifiers. The RF-trained model selects all importance of all features [31]. As shown in Fig. 3, we present the Brouta feature and how it selects the most crucial feature to add to classifier [32]. The algorithm of Boruta consists of the following steps:

1. Expanding the information system by adding copies of all variables,$(x_t^{'})$ for a particular input vector,x_v to add randomness and to eliminate the correlations between duplicate predictors and targets (y_t), for a group of discrete inputs ($X_t \in R^n$), T and target variable ($y_t \in R$) with several inputs (n) and $t = 1, 2, \ldots T$.
2. The added attributes are shuffled to remove their correlations with the response.
3. Running a classifier RF on the expanded information system with the target (y_t) forecast the duplicated $(x_t^{'})$ and actual (x_t) inputs.
4. The variance importance measures are used, i.e., permutation importance or mean decrease accuracy (MDA) for each input x_t and respective shadow input $(x_t^{'})$ overall trees (m_{tree}) by Eq. (4). $I(\cdot)$ is the indicator function. OOB is derived as Out-of-Bag and it is the prediction error of each of the training samples based on bootstrap aggregation. $(y_t = f(x_t))$ are predicted values before permuting; and $(y_t = f(x_t^n))$ is defined as the predicted values after permuting.

$$MDA = \frac{1}{m_{tree}} \sum_{m=1}^{m_{tree}} \frac{\sum_{t \in OOB} I(y_t = f(x_t)) - \sum_{t \in OOB} I(y_t = f(x_t^n))}{|OOB|} \tag{4}$$

5. The calculated Z scores are collected by using Eq. (5). Here, SD is the standard deviation of accuracy losses.

$$Zscore = \frac{MDA}{SD} \tag{5}$$

6. The maximum Z score is found between shadow attributes (MZSA). Then, a score is assigned for each attribute, which is better than MZSA.
7. For each characteristic of the indeterminate significance of attributes, a two-sided equality test with MZSA is run.
8. Features of much less significance than MZSA are considered ”insignificant” and permanently removed from the information system.
9. The features of much higher importance than MZSA are considered ”important”.
10. All shadow attributes are removed.
11. The procedure is repeated from 1 to 10 until all attributes are assigned importance, or the algorithm has reached a predetermined limit for RF runs.

3.3.2. *Information gain (IG)*

IG method depends on the concept of entropy. In filter strategies, it is used as an evaluator of feature relevance that rates features individually, and it has the advantage of being quick [33]. IG can identify the needed features from each class. It is driven by entropy using Eq. (6). Let D (A_1, A_2, …, A_n, C), $n \geqslant 1$, be an $n+1$ attribute dataset, where C is an attribute of the class. Set m be the number of values of distinct classes. The class distribution of entropy in D is represented by Entropy(D), is defined by Eq. (6) [34].

$$Entropy(D) = \sum_{i=1}^{m} p_i * \log_2 * p_i \tag{6}$$

where p_i is the probability that an arbitrary instance in D belongs to class c_i.

3.4. *Diagnosis by GBT*

In all studies for classifying diseases and detecting hidden diseases' characteristics, ML is widely used. Also, the overall performance improves by combining different ML techniques. This paper main objective is to apply a comprehensive ML-based system for determining genetic biomarkers associated with early AD stages. In this section, the proposed work is applied for early detection of AD, and a comparison with other ML techniques, such as NB, RF, SVM, and KNN, is made. In Fig. 4, a Manhattan plot is presented. The most commonly used scatter plot to represent data is the Manhattan plot, which is preferable to use with many data points, many non-zero amplitudes, and a distribution of higher-magnitude values. The plot is usually used in GWAS to show significant SNPs. Each point represents a genetic variant. The X-axis shows the position on a chromosome, whereas the Y-axis tells how much it is associated with a trait. The black and gray colors, which are used in

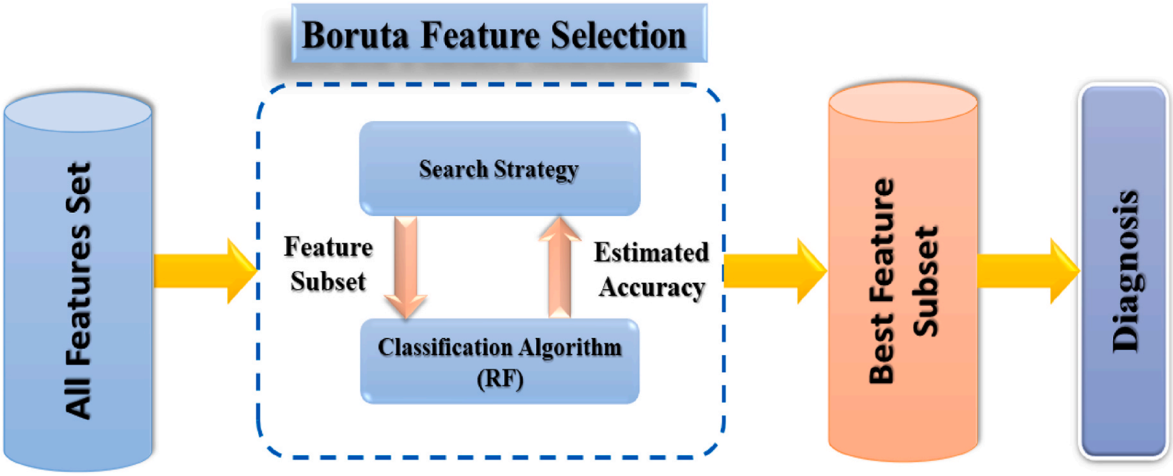

Fig. 3. The feature selection process using Boruta technique.

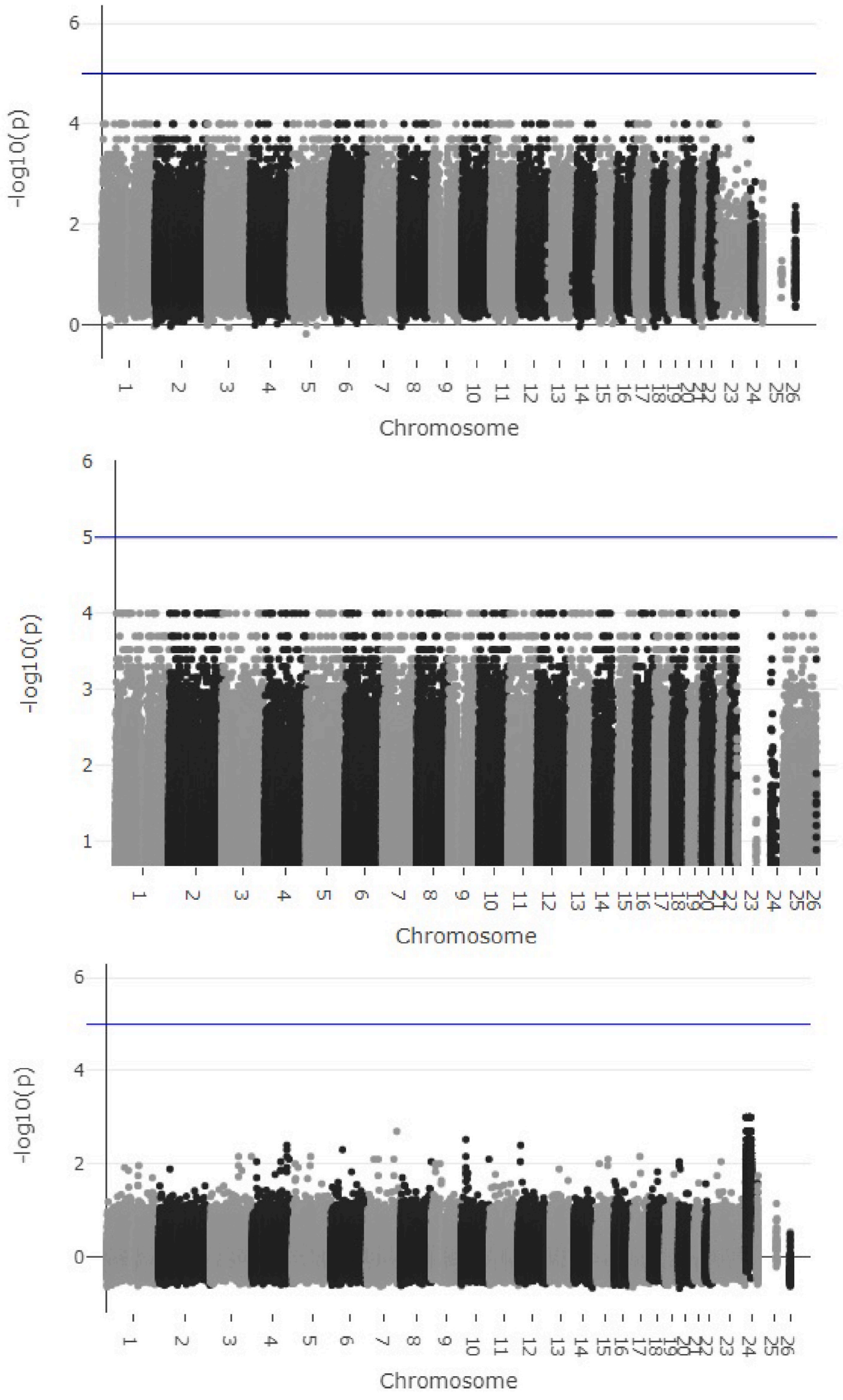

Fig. 4. Examples of Manhattan plots: (a) The top plot for normal case, (b) The middle plot for MCI, and (c) The bottom plot for AD.

the Manhattan plot, are utilized to distinguish the adjacent chromosomes from each other. They are used to illustrate the data size and boundaries of each chromosome.

As we know, each human carry 23 chromosomes, which are included in the standard ADNI database. Chromosome number 23 can take one of the following values: X, Y, XY, and XX. These values are represented in the X-axis of the Manhattan plot with the numbers 23, 24, 25, and 26, respectively. As shown in Fig. 4, a Manhattan plot is presented for three cases: normal, MCI, and AD. CLU, ABCA 7, APOE, BINI, CRI, CD2AP, and CD33 are the most highly associated genes with AD, which were identified as the most significant SNPs. These SNPs are regarded as critical biomarkers for the disease. A number of common SNPs are obtained as a significant SNPs for early detection of AD as follows: rs512941, rs2074451, rs429358, rs1595816, rs17014396, rs-9296559, rs2007332, and rs204 I 992.

3.4.1. *Gradient boosting*

Gradient boosting is an ensemble learning method, which iteratively extremely adds basic models. Each additional basic model further reduces the gradient of the specified loss (error) function [35]. x refers to the feature vector, and y refers to the class label. Given some training samples $\{x_i, y_i\}_{1=1}^{N}$,the main aim is to obtain a function $F^{*}(x)$ that can transfer x into y, like the expected value of a given loss function $L(y; F(x))$ is minimized over the joint distribution of $\{x, y\}$ values. The loss function measures the deviation between the real value y and the predicted value $\hat{y}$. The ”additive” expansion is expressed to approximate the function by Eq. (7).

$$F^{*}(x) = \arg\min_{F(x)} E_{y,x}L(y, F(x)) = \arg\min_{F(x)} E_x[E_yL(y, F(x))|x] \tag{7}$$

$$F(x; P) = \sum_{m=1}^{M} \beta_m h(x; \gamma m) \tag{8}$$

where $P = \{\beta_m, \gamma_m\}_{m=0}^{M}$ the function $h(x; \gamma)$, known as ’base learner’, is always a simple function of x with parameters $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_M\}$. If $F(x)$ is estimated in a nonparametric manner, then the task will become more complicated. So, the function optimization problem must therefore be mapped to the parameter optimization problem by choosing a model $F(x; P)$, which can be set by P. A ”greedy-stagewise” approach is a typical parameter optimization method. $\{\beta_m, \gamma_m\}$ is optimized after all the $\{\beta_i, \gamma_i\}(i = 0, 1, \ldots, m - 1)$ are optimized. This process can be formulated by Eq. (9). Table 2 lists the main steps of the gradient boosting algorithm for SNPs.

$$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=0}^{M} L(y_i, F_{m-1}(x_i)) + \beta h(x_i, \gamma) \tag{9}$$

$$F_m = F_{m-1} + \beta_m h(x, \gamma_m) \tag{10}$$

4. Experimental results

This section consists of two subsections: the evaluation metrics and results. The used performance measures are detailed in the evaluation metrics subsection. In the results subsection, the proposed system results are presented as well as the overall proposed system on the benchmark dataset. Then, an overall comparison is presented between the proposed system with state-of-the-art ML techniques. In addition, an analytical comparison in the results subsection between the proposed system and different ML techniques is provided.

4.1. Evaluation metrics

The used performance metrics were accuracy (ACC), precision (Prec), sensitivity (Sens), the area under the curve (AUC), and Disc similarity coefficient (DSC) in the classification stage. ACC is the ratio of correctly predicted populations to the total populations. Prec is the ratio of correctly predicted positive populations to the total predicted positive populations. Sens is the ratio of correctly predicted positive populations to the actual positive populations. The receiver operating characteristic (ROC) curve is another measure of the performance of an ML classifier model. FP is the ratio of false predictive or incorrect positive predictions. FN is the ratio of incorrect negative predictions. The ROC is a probability curve, while the AUC represents the capability of the model to distinguish among classes. The ROC curve is constructed by plotting the true positive rate versus the false positive rate. DSC is a statistical tool that measures the similarity between two sets of data, as shown in Table 3 [36–39].

Table 2
Gradient boosting algorithm.

```
Algorithm: Gradient boosting for SNPs
Input phase:
All SNPs set x;
The iterative steps, M;
Output phase:
The function of final classification $F_m(x)$;
initialize $F_0(x) = \arg\min_p \sum_{i=1}^{N} L(y_i, p)$;
for m = 1 to M do
calculation of the negative gradient.
$\widetilde{y}_m = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$
Fit a model.
$\alpha_m = \arg\min_{\gamma,\beta} \sum_{i=1}^{N} [\widetilde{y} - \beta h(x_i; \alpha_m)]^2$
select a gradient descent step size as
$Pm = \arg\min_P \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i)) + ph(x_i, \alpha)$
Update the estimation of F(x)
$F_m(x) = F(m-1)(x) + pmh(x; \alpha m)$
End for
Return $F_m(x)$:
```

5. Results

This subsection presents a comprehensive framework for early detection of AD and detecting the most significant genes based on SNPs analysis. In this work, six ML techniques: GBT, NB, RF, SVM with linear kernel, SVM with RBF, and KNN are used. Also, the two feature selections, which are Boruta and IG, are used. First, the dataset is shuffled in case of using k-fold cross-validation, so the order of the inputs and outputs is entirely random. This step is performed to ensure that the inputs are not biased in any way. The division of the dataset into k parts of equal sizes is done subsequently. To evaluate the performance of classifiers with the implemented models, the k-fold cross-validation is applied. The dataset is divided into approximately ten equal groups referred to as k = 10. There is a similar percentage for each group of people who contracted the disease. Classifier designing depends on 9/10 of the datasets. The rest of the data was regarded as a test set to evaluate the classifier’s performance. For each test group, this operation is then repeated ten times [40]. All parts of the proposed system are implemented by using the R language. The proposed system is developed on a machine with an i7/2.6 GHz processor and 8 GB RAM. The code is uploaded as a supplemental file with this article.

In Table 4, the hyperparameter optimization of the proposed techniques is presented. Boruta feature selection is a warper feature selection. It will generate all possible subsets of the dataset. Then, the classification algorithm is performed to induce classifiers from the features in each subset. As shown in Table 6, Boruta feature selection is used with six ML classifiers. New strategies for detecting, treating, and preventing the disease can be followed to discover SNPs biomarkers associated with AD. For using the ADNI-1 dataset, 26,734 SNPs are selected for the feature selection step. The total number of selected SNPs is 75,772 in the case of using the WGS dataset. 2-fold and 10-fold cross-validation techniques are applied. The results show that GBT is superior to NB, RF, SVM with RBF, and KNN with 98% accuracy in case 2-fold cross-validation. Also, results show that GBT is superior to the five

Table 3
The used performance evaluation metrics.

Metrics	Description	Formula
accuracy	This is a relation between the sum of TP and TN divided by the total sum of the population	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity, Recall	This is a relation between TP divided by the total sum of TP, FN	$Sens = \frac{TP}{TP + FN}$
Specificity	This is a relation between TN divided by the total sum of TN, FP	$Spec = \frac{TN}{TN + FP}$
AUC	This metric is used to measure the average area under ROC	$TPR = \frac{TP}{TP + FN}$ $FPR = \frac{FP}{FP + TN}$
DSC	This is a relation between TP divided by the total sum of TP, FN, FP	$DSC = \frac{2 * TP}{2 * TP + FP + FN}$

Table 4
The hyperparameters of the used techniques.

Method	Hyperparameter
Boruta	pValue = 0.01, where pValue: is the confidence level. Default value of maxRuns = 100, maxRuns is the maximal number of importance source runs. Classification algorithm is RF.
SVM	Use two kernels are linear kernel and RBF. method = cross validation with number = 2,10
RF	default = 100, where the number of trees in the forest. method = cross validation with number = 2,10
GBT	n.trees = 100(number of trees). learning rate = 0.001, Cross validation. folds = 2,10

used ML techniques with 99.06% accuracy in the case of 10-fold. Table 5 represents the top selected SNPs and top candidate genes related to the early detection of the disease.

In Table 7, six classifier ML techniques and IG feature selection are used. IG is a filter feature selection. An evaluator and a ranker are used to rank all features for a more precise dataset. The attributes are arranged according to the rank. Then, a comparison between the six classifiers is made. The results show that GBT is superior to the other five classifiers used in 2-fold and 10-fold cross-validation techniques. Also, the results show the ROC curves for five different classifiers, as shown in Fig. 5. Finally, the Boruta feature selection is superior to IG.

Fig. 5 shows the ROC curves for the proposed framework and other tested classifiers. The proposed system in cases of using Boruta feature selection has a good ROC curve. If the AUC is close or equal to 1, it is a good classifier. On the other hand, it means that GBT is superior to other classifiers as it manifests the highest values of the ROC curve. As shown in Table 6 and Fig. 5, the Boruta feature selection has the highest accuracy within the classifiers. Accordingly, the wrapper feature selection is better than the filter feature selection.

In Table 8, our framework is applied to the WGS dataset. 2-fold and 10-fold cross-validation techniques and the five measures, which are ACC, Spec, Sens, AUC, and DSC, are used to validate the work. The results show that GBT has an ACC of 99%, Spec is 97.4%, Sens is 99.27%, AUC is 0.98, and DSC is 98.08%. These results demonstrate the effectiveness of using the applied ML techniques for early detection or early-stage AD detection. Therefore, ML is a good tool that can help doctors diagnose the disease early.

Table 5
The top selected SNPs among the top candidate genes.

CHR	GENE	SNPs
8	CLU	rs512941
19	**ABCA 7**	**rs2074451**
19	APOE	rs429358
2	**BINI**	**rs1595816**
1	CRI	rs2660635
6	**CD2AP**	**rs9296559** **rs17014396**
19	CD33	rs2007332 rs20n561

Table 6
The performance evaluation of our framework system using some state-of-the-art classification techniques and Boruta feature selection approach on the ADNI-1 dataset.

Boruta Feature selection							
Cross validation	Metrics	GBT	NB	RF	SVM with linear	SVM with RBF	KNN
k = 2	ACC.	98	97.89	97.89	96	96.67	93
	Spec.	98.78	98.5	97.1	97.68	98.65	95
	Sens.	99.02	99.56	99.03	96	97.87	96.43
	AUC	.97	0.96	0.961	0.95	0.95	0.923
	DSC	96.54	95.7	95.02	93.84	92.56	90.02
k = 10	ACC.	99.06	98.1	97.97	95.88	96.98	93
	Spec.	99	98.04	97.15	97.68	98	96
	Sens.	98.45	99.56	99.03	96.26	98.43	94.46
	AUC	.98	0.97	0.96	0.94	0.9587	0.928
	DSC	97.32	96.01	94.89	93.67	94.05	92

Table 7
The performance evaluation of our framework system using some state-of-the-art classification techniques and IG feature selection approach on the ADNI-1 dataset.

IG Feature selection							
Cross validation	Metrics	GBT	NB	RF	SVM with linear	SVM with RBF	KNN
k = 2	ACC.	94.64	94.2	93	92	92	83
	Spec.	95	96	97.6	97	97	88.53
	Sens.	98.03	97.98	94	95.8	95.8	82
	AUC	.937	0.93	.91	0.91	0.91	0.82
	DSC	91.5	90.34	88.05	89.25	90.04	80
k = 10	ACC.	94.87	94.2	93.5	91.06	92	85
	Spec.	96.90	96	96.7	96	97	88.01
	Sens.	98	97.98	95.88	95	96	90.98
	AUC	.94	0.93	0.929	0.91	0.91	0.84
	DSC	91.86	90.34	90.88	90.34	88.45	84.29

Table 9 shows the comparison of our framework system with other-state-of-the-art techniques with respect to time. Also, we used two deep learning techniques: multilayer perceptron (MLP) and recurrent neural network (RNN), to compare with our proposed framework. MLP and RNN take the longest time to train the model. The MLP achieved an accuracy of 91.57% with the epoch number = 25 with two hidden layers. In the case of increasing the epoch number, the accuracy increased, but the overfitting between the training and validation increased. Also, RNN is applied with epoch number = 49 with two hidden layers and achieved an accuracy of 94.50%. Finally, we can conclude that RNN is more efficient than MLP. Also, we can conclude that ML and deep learning are more efficient for early detection of AD using SNPs. Also, missing data causes a variety of issues, which can be summarized in the following points:

1. A lack of data reduces statistical power, which is the likelihood that the test would reject the null hypothesis when it is wrong.
2. Lost data might lead to bias in a parameter estimate.
3. It has the potential to impair the representativeness of the samples.
4. It may complicate the study's analysis. Each of these distortions can affect the validity of the trials and lead to incorrect findings.
5. Two imputation methods, mean and median, were performed to handle the missing data.

As shown in Tables 6–8, we use mean to handle missing data and outliers. In Table 9, we use the median to handle missing data and outliers. The median performs more accurately than the mean.

Our work's primary concern or novelty is detecting the disease in an early stage and identifying the most significant genes and SNPs that cause the disease, as shown in Table 10. Our framework has the highest accuracy in comparison to other systems.

6. Discussion

This paper proposes a system for the early detection of AD and the classification process. The proposed framework utilizes two FS techniques with ML. As shown in Tables 6 and 7, the Boruta FS is superior to IG because Boruta deals with all features in contrast to IG search for the subsets of features. As shown in Tables 6 and 7, results revealed that the proposed system, NB, RF, SVM, SVM with RBF, and KNN learning

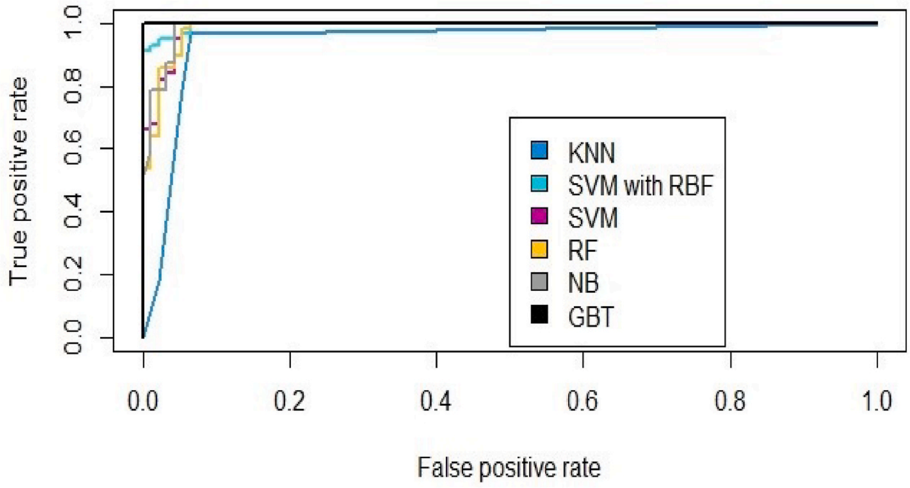

Fig. 5. The ROC for ML classification techniques.

Table 8
The performance evaluation of our framework system using some state-of-the-art classification techniques and Boruta feature selection approach on the WGS dataset.

Cross validation	Metrics	GBT	NB	RF	SVM with linear	SVM with RBF	KNN
k = 2	ACC.	98.56	98.02	96.9	97	97.8	93.3
	Spec.	97	96.45	95.54	99.87	99.06	96
	Sens.	99.37	99	99.03	94.28	97	91.87
	AUC	.977	0.97	0.94	0.94	0.95	0.92
	DSC	98	94.7	93.76	95.1	93	89
k = 10	ACC.	99	98.87	97	96.91	98	93.26
	Spec.	97.4	99	98	98.05	98	90
	Sens.	99.27	97.05	95.8	94.2	96.95	97.97
	AUC	.982	0.98	0.963	0.95	0.96	0.92
	DSC	98.08	93.98	93.98	93	94	91

Table 9
The classification accuracy using median imputation and computation time for used techniques.

Methods	Accuracy(%)	Time(minutes)
GBT	99.23	66
NB	98.15	51
RF	97.99	56
SVM	95.98	40
SVM-RBF	97.04	43
KNN	93.45	80
RNN	94.50	120
MLP	91.57	100

Table 10
The comparison between the proposed system and others systems.

Authors	Methods	Accuracy(%)
Abd El Hamid et al. [19]	SVM	76.70
Bringas et al. [41]	CNN	90.91
Shahbaz et al. [42]	GLM	88.24
Proposed System	GBT	99.23

algorithms scored an overall ACC of 99.06%, 98.1%, 97.97%, 95.88%, 96.67%, and 93%, respectively in case of using Boruta. In case of using IG feature selection, the scored accuracies are as follows: 94.87%, 94.2%, 93.5%,91.06%,92%, and 85%, respectively. The results show that the classification techniques are favorable for detecting AD in the early stages. As shown in the results, the GBT is superior to all classifiers.

From Tables 6 and 7 in the whole-genome approach (AD–NI–1), it was observed that there is convergence in the results. These results demonstrated the effectiveness of using the applied ML techniques for identifying significant SNPs associated with the disease with acceptable ACC. In WGS, it was shown that there is convergence in the results presented in Fig. 5. The highest classification ACC was achieved by GBT, which equals 99%. However, SVM is trained using linear kernel and RBF.

The work of Abd El Hamid et al. [19] holds similarities to the proposed work. The main goal of their work is to determine the most critical forms of SNP associated with AD. It decreases the computational complexity of ML techniques rather than dealing with all features and obtaining a high classification performance. They also use different techniques for feature selection to select the best subset features and eliminate unimportant and redundant features.

Shahbaz et al. [42] presented AD classification by using ML techniques to classify AD. The dataset is split into a training dataset with 70% partition and the remaining percentage testing. However, the model cannot be trained with unbalanced and insufficient data for all disease classes. In the proposed system, unbalanced data is taken into account using k-fold cross-validation. 2 and 10 folds cross-validation techniques are used. Cross-validation is a very useful tool as it supports better use of data. Moreover, it provides much more information about the performance of algorithms. The preprocessing phase is also another point of great importance. The most significant advantage of preprocessing in ML is to improve the generalizability of the model.

Sherif et al. [6] identified several significant polymorphisms associated with AD in the APOE, CR1, CD33, CLU, PICALM, and ABCA7 genes. In our framework, we identified CLU, ABCA 7, APOE, BINI, CRI, CD2AP, and CD33 are the most highly associated SNPs with AD. On the other hand, our framework achieved the highest accuracy of classification.

Some of the SNPs were linked with AD risk in studies that have previously been discovered, including rs113464261, rs769449, rs73504429, APOE112, rs4844609, rs4732729, rs9331942, rs610932, rs7530069, rs114506298, rs7929589, and rs611267. Other additional SNPs were discovered to be highly related to AD. rs512941, rs2074451, rs1595816, rs17014396, rs9296559, and rs2007332 are the SNPs involved.

All studies should take into consideration the preprocessing phase. FS is recommended to decrease the number of highly correlated SNPs. Highly correlated SNPs make it challenging to select the true disease-causing variant. FS can be used in both supervised and unsupervised learning. However, this paper concentrates on the problem of supervised learning (classification), where class labels are predefined in advance. It is crucial to decrease the dimensions by methods such as FS because high-dimensional feature vectors of microarray data often lead to high computational costs. In addition, the risk of overfitting due to the extended classification time becomes a high probability of the overfitting [2,5]. The dataset should be normalized so the data is mapped in a specific range and so there are no missing values of the data to avoid misclassification. Then, the feature is selected, and as a result, the main genes are obtained. The classification is then performed, and the outputs are interpreted to have the required biological information. Accordingly, the feature should only be selected once then the evaluation for different classifiers can be performed. The main drawback of filtering methods is that they neglect interaction with the classifier, and each feature is considered independently. It also neglects features' dependency on each other.

In contrast to simple wrappers interacting with the classifier, models are characterized by dependency, good classification accuracy, and computational cost reduction. The only main drawback is being computationally intensive. In conclusion, feature selection has great importance in classification for the following reasons: it reduces the effects of the curse of dimensionality, helps in model learning, reduces the cost of computation, and helps achieve reasonable accuracy.

7. Conclusion

The main objective of the recent studies is to identify genetic biomarkers for complicated diseases, including AD. It is a critical stage for identifying genes involved in AD development in molecular diagnostics. This study is based on genetic data of ADNI-1 and WGS datasets with SNPs tested using many phenotypes. In ADNI-1, results revealed that the proposed system scored 99.06%,98.1%, 97.97%, 95.88%, 96.67%, and 93% for NB, RF, SVM, SVM with RBF, and KNN learning, respectively, in case of using Boruta FS. Using information gain FS, the proposed system achieved accuracy equals 94.87%, 94.2%, 93.5%,91.06%,92%, and 85% for NB, RF, SVM, SVM with RBF, and KNN learning, respectively. In the case of WGS data, results showed that the proposed system achieved an overall accuracy of 99%, 98.87%, 97%, 96.91%, 98%, and 93.26% for NB, RF, SVM, SVM with RBF, and KNN learning, respectively. CLU, ABCA 7, APOE, BINI, CRI, CD2AP, and CD33 are the most highly associated SNPs with AD, which were identified. These SNPs are regarded as critical biomarkers for the disease. Consequently, the proposed system is proved to help improve medical diagnosis methods and identify the causes of the disease at its earliest stages. The combination of genetic data and medical images will be processed for future works to obtain a comprehensive early detection system for AD.

Declaration of competing interest

None Declared.

References

[1] Aziz M. Mezlini, Anna Goldenberg, Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases, PLoS Comput. Biol. 13 (10) (2017), e1005580.

[2] Andrew J. Saykin, Li Shen, Tatiana M Foroud, Steven G. Potkin, Shanker Swaminathan, Sungeun Kim, Shannon L. Risacher, Kwangsik Nho, Matthew J. Huentelman, David W. Craig, et al., Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans, Alzheimer's Dementia 6 (3) (2010) 265–273.

[3] A El-Gamal Fatma El-Zahraa, Mohammed M. Elmogy, Ashraf Khalil, Mohammed Ghazal, Soliman Hassan, Atwan Ahmed, Robert Keynton, Gregory N. Barnes, S El-Baz Ayman, A significant regional-based diagnosis system for early detection of alzheimer's disease using smri scans, in: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2018, pp. 407–412.

[4] Priyanka Nakka, Benjamin J. Raphael, Sohini Ramachandran, Gene and network analysis of common variants reveals novel associations in multiple complex diseases, Genetics 204 (2) (2016) 783–798.

[5] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[6] Fayroz F. Sherif, Nourhan Zayed, Mahmoud Fakhr, Discovering alzheimer genetic biomarkers using bayesian networks, Adv. Bioinformatic. (2015) 1–8.

[7] Yvan Saeys, Inaki Inza, Pedro Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[8] Chunming Xu, Scott A. Jackson, Machine learning and complex biological data, Genome Biol. 1–4 (2019).

[9] W Libbrecht Maxwell, Stafford Noble William, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332.

[10] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Torkamani Ali, Amalio Telenti, A primer on deep learning in genomics, Nat. Genet. 51 (1) (2019) 12–18.

[11] Mirza Jawad Ul Hasnain, Muhammad Shoaib, Salman Qadri, Bakhtawar Afzal, Tehreem Anwar, Syed Hassan Abbas, Amina Sarwar, Hafiz Muhammad Talha Malik, Muhammad Tariq Pervez, Computational analysis of functional single nucleotide polymorphisms associated with slc26a4 gene, PLoS One 15 (1) (2020), e0225368.

[12] Walid Korani, Josh P. Clevenger, Ye Chu, Peggy Ozias-Akins, Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants, Plant Genome 12 (1) (2019) 1–10.

[13] Francis S. Collins, Lisa D. Brooks, Aravinda Chakravarti, A DNA polymorphism discovery resource for research on human genetic variation, Genome Res. 8 (12) (1998) 1229–1231.

[14] Widi Astuti, et al., Support vector machine and principal component analysis for microarray data classification, in: Journal of Physics Conference Series, vol. 971, 2018, 012003.

[15] Marjane Khodatars, Afshin Shoeibi, Delaram Sadeghi, Navid Ghaasemi, Mahboobeh Jafari, Parisa Moridian, Khadem Ali, Roohallah Alizadehsani, Assef Zare, Yinan Kong, et al., Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review, Comput. Biol. Med. 139 (2021) 104949.

[16] S.R. Bhagya Shree, H.S. Sheshadri, Diagnosis of alzheimer's disease using naive bayesian classifier, Neural Comput. Appl. 29 (1) (2018) 123–132.

[17] A. Suppers, A. J van Gool, H.J.C.T. Wessels, Integrated chemometrics and statistics to drive successful proteomics biomarker discovery, Proteomes 6 (2) (2018) 20.

[18] Mark Nnh Mikhail, Y Sayed Ahmed, Mai S. Mabrouk, Ayman M. Eldeib, Investigation of genome-wide association SNPs and alzheimer's disease, Am. J. Biomed. Eng. 10 (1) (2020) 1–8.

[19] Marwa Mostafa Abd El Hamid, Mai S. Mabrouk, Yasser MK. Omar, Developing an early predictive system for identifying genetic biomarkers associated to alzheimer's disease using machine learning techniques, Biomed. Eng.: Appl. Basis. Commun. 31 (5) (2019) 1950040.

[20] Marwa Mostafa Abd El Hamid, Yasser MK. Omar, Mai S. Mabrouk, Identifying genetic biomarkers associated to alzheimer's disease using support vector machine, in: 2016 8th Cairo International Biomedical Engineering Conference (CIBEC), IEEE, 2016, pp. 5–9.

[21] Matt Spencer, Nicole Takahashi, Sounak Chakraborty, Judith Miles, Shyu Chi-Ren, Heritable genotype contrast mining reveals novel gene associations specific to autism subgroups, J. Biomed. Inf. 77 (2018) 50–61.

[22] Aicha Boutorh, Guessoum Ahmed, Classification of SNPs for breast cancer diagnosis using neural-network-based association rules, in: 2015 12th International Symposium on Programming and Systems (ISPS), IEEE, 2015, pp. 1–9.

[23] Barath Narayanan Narayanan, Russell C. Hardie, Temesguen M. Kebede, Performance analysis of feature selection techniques for support vector machine and its application for lung nodule detection, in: NAECON 2018-IEEE National Aerospace and Electronics Conference, IEEE, 2018, pp. 262–266.

[24] Yang Hu, Tianyi Zhao, Tianyi Zang, Ying Zhang, Liang Cheng, Identification of alzheimer's disease-related genes based on data integration method, Front. Genet. 9 (2019) 703.

[25] Sumit Mukherjee, Thanneer M. Perumal, Kenneth Daily, Solveig K. Sieberts, Larsson Omberg, Christoph Preuss, Gregory W. Carter, Lara M. Mangravite, Benjamin A. Logsdon, Identifying and ranking potential driver genes of alzheimer's disease using multiview evidence aggregation, Bioinformatics 35 (14) (2019) i568–i576.

[26] ADNI dataset:. http://adni.loni.usc.edu/. (Accessed 28 December 2021).

[27] Sergio Campos, Luis Pizarro, Carlos Valle, Katherine R. Gray, Daniel Rueckert, Héctor Allende, Evaluating imputation techniques for missing data in ADNI: a patient classification study, in: Iberoamerican Congress on Pattern Recognition, Springer, 2015, pp. 3–10.

[28] Roberto M. Cesar Jr., Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2010, pp. 477–483.

[29] Maisa Daoud, Michael Mayo, A survey of neural network-based cancer prediction models from microarray data, Artif. Intell. Med. 97 (2019) 204–214.

[30] S. Karthik, M. Sudha, A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases, Int. J. Eng. Adv. Technol. 8 (2) (2018) 182–191.

[31] Pi Guo, Youxi Luo, Guoqin Mai, Ming Zhang, Guoqing Wang, Miaomiao Zhao, Liming Gao, Li Fan, Fengfeng Zhou, Gene expression profile based classification models of psoriasis, Genomics 103 (1) (2014) 48–55.

[32] Miron B. Kursa, Witold R. Rudnicki, et al., Feature selection with the boruta package, J. Stat. Software 36 (11) (2010) 1–13.

[33] Jaesung Lee, Dae-Won Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recogn. Lett. 34 (3) (2013) 349–357.

[34] Nermeen A. Shaltout, Mahmoud El-Hefnawi, Rafea Ahmed, Moustafa Ahmed, M. El-Hefnawi, Information gain as a feature selection method for the efficient classification of influenza based on viral hosts, Proc. World Congress Eng. 1 (2014) 625–631.

[35] Longquan Jiang, Bo Zhang, Qin Ni, Xuan Sun, Pingping Dong, Prediction of SNP sequences via gini impurity based gradient boosting method, IEEE Access 7 (2019) 12647–12657.

[36] Indu Jain, Vinod Kumar Jain, Renu Jain, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification, Appl. Soft Comput. 62 (2018) 203–215.

[37] Rabindra Kumar Singh, M. Sivabalakrishnan, Feature selection of gene expression data for cancer classification: a review, Procedia Comput. Sci. 50 (2015) 52–57.

[38] Suman Raj, Sarfaraz Masood, Analysis and detection of autism spectrum disorder using machine learning techniques, Procedia Comput. Sci. 167 (2020) 994–1004.

[39] Kwetishe Joro Danjuma, Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients, IJCS Int. J. Computer Sci. Iss. 12 (2015) 1–11.

[40] Peter S. Kristensen, Jahoor Ahmed, Jeppe R. Andersen, Fabio Cericola, Jihad Orabi, Luc L. Janss, Just Jensen, Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines, Front. Plant Sci. 9 (2018) 69.

[41] Santos Bringas, Sergio Salomón, Rafael Duque, Carmen Lage, José Luis Montaña, Alzheimer's disease stage identification using deep learning models, J. Biomed. Inf. 109 (2020) 103514.

[42] Muhammad Shahbaz, Shahzad Ali, Guergachi Aziz, Aneeta Niazi, Amina Umer, Classification of alzheimer's disease using machine learning techniques, in: DATA, 2019, pp. 296–303.