IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Explainable Machine Learning Model for Alzheimer Detection Using Genetic Data: A Genome-Wide Association Study Approach

TAREK KHATER[1, 2], SAM ANSARI[2], ABBAS S. ALATRANY[3, 4], HAYA ALASKAR[5], SOLIMAN MAHMOUD[2, 6], (SENIOR MEMBER, IEEE), AYAD TURKY[7], HISSAM TAWFIK[2], (SENIOR MEMBER, IEEE), EQAB ALMAJALI[2],(MEMBER, IEEE), AND ABIR HUSSAIN[2, 3], (SENIOR MEMBER, IEEE).

[1]Biomedical Engineering Department, Khalifa University, Abu Dhabi, United Arab Emirates
[2]Electrical Engineering Department, University of Sharjah, Sharjah, United Arab Emirates
[3]School of Computer Science and Mathematics, Faculty of Engineering, Liverpool John Moores University, Liverpool, L3 3AF, United Kingdom
[4]University of Information Technology and Communications, Baghdad, Iraq
[5]Department of Computer Sciences, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[6]University of Khorfakkan, Khorfakkan, United Arab Emirates
[7]Department of Computer Science, College of Computing and Informatics, University of Sharjah, Sharjah, United Arab Emirates

Corresponding author: Abir Hussain (abir.hussain@sharjah.ac.ae; a.hussain@ljmu.ac.uk)

ABSTRACT Recent research has revealed that using machine learning systems for the analysis of genetic data could reliably detect Alzheimer's disease. The interpretability of these models, however, has been a challenge, as they frequently provided little insight into the features that contribute to their predictions. Explainable machine learning has been presented as a solution to this problem since it enables the identification of significant attributes and gives a clearer method of making predictions. In this study, Genome-Wide Association Studies were used to recognize genetic variants associated with Alzheimer's disease, utilizing the Alzheimer's Disease Neuroimaging Initiative dataset and quality control methods to ensure the validity and reliability of the findings. The results indicate strong connections between certain genetic variations and Alzheimer's disease, highlighting the potential of Genome-Wide Association Studies as a valuable tool for identifying and predicting this disease. After studying and analyzing the genetic data, machine learning algorithms are utilized to train a model to detect Alzheimer. The Support Vector Machine achieved 89% accuracy as the best-performing model. Explainable machine learning has the potential to increase the accuracy and interpretability of Alzheimer's disease detection models, giving significant insights for both academics and physicians. The explanation of the support vector machine model reveals that rs4821510 is the most important SNP in detecting AD. On top of that, the SHAP method shows that rs429358 is an indication for Alzheimer's disease and rs4821510 presents in the healthy ones. These findings suggest that explainable machine learning can play an important role in accurately detecting Alzheimer's disease and identifying critical genetic markers associated with the disease.

INDEX TERMS Alzheimer, artificial intelligence, GWAS, quality control, XAI.

I. INTRODUCTION

Alzheimer's disease is a progressive, neurodegenerative brain disorder that affects memory, thinking skills, and the capability to carry out daily activities [1]. The disease is recognized by a gradual loss of memory and basic life skills such as eating, bathing, and talking [2]. Symptoms of Alzheimer's disease include memory loss, paranoia, depression, anger, aggression, anxiety, apathy, loneliness, and psychosis [2].

Alzheimer's disease is the most common cause of dementia, representing around 70% of all cases [3]. The disease usually affects individuals over the age of 65, with symptoms appearing in their mid-60s [3]. However, a rare form of the disease, known as early-onset Alzheimer's, can occur in individuals between their 30s and mid-60s [3]. Alzheimer's

disease affects an estimated 6.5 million Americans aged 65 and older today [4]. Alzheimer's disease and dementia cases are on the rise in the United Arab Emirates (UAE) as well. It is forecasted that the UAE will see a significant increase in dementia cases, with a predicted 1,795 percent rise by the year 2050 [5]. This is one of the second-highest percentage increases in dementia cases globally. The neighboring Gulf countries, such as Qatar and Bahrain, are also expected to experience similar trends.

If no medical breakthroughs are made to prevent, cease, or cure Alzheimer's disease, this figure might rise to 13.8 million by 2060 [4]. A recent survey conducted by the American Alzheimer's Association [6] identified many impediments to consumers' awareness of Mild Cognitive Impairment (MCI), a condition that may raise the chance of acquiring Alzheimer's disease. According to the report, Americans are unaware of MCI and are hesitant to seek medical help. The report also projects that total expenses for healthcare, long-term care, and hospice services for those 65 and older with dementia in 2022 will be $321 billion. As the disease progresses, it can lead to moderate to severe cognitive impairment, affecting areas of the brain that control languages, reasoning, conscious thought, and sensory processing, such as the ability to correctly detect sounds and smells [7]. Memory loss and confusion also worsen, and people with Alzheimer's disease may have difficulty recognizing family and friends [7].

Understanding the early signs and symptoms of Alzheimer's disease can help with early diagnosis and treatment. Ongoing research is aimed at identifying the underlying causes of the disease and developing effective therapies to slow or stop its progression with the aid of Artificial Intelligence (AI) [8]. AI is used to detect Genetic Alzheimer's Disease for a variety of reasons. One important reason is drug delivery, where AI plays a crucial role in repurposing existing drugs for AD treatment [9]. It can quickly analyze large amounts of data, such as transcriptomics, molecular structures, and clinical databases, to predict drug repurposing. This offers a fast and cost-effective way to develop drugs [9]. In addition, AI contributes significantly to genetic research on AD. It helps with the diagnosis, prognosis, and analysis of genetic data related to AD. This includes studying genetic variation, gene expression profiles, gene-gene interactions, and utilizing knowledge bases for genetic analysis [10].

For instance, Rs429358 and rs4420638 are two common polymorphisms located within the APOE gene, which encodes apolipoprotein E (APOE), a protein involved in lipid transport and metabolism in the brain. Numerous studies have identified these variants as strong genetic risk factors for AD. Specifically, the ϵ_4 allele of rs429358 and the ϵ_2 allele of rs4420638 have been consistently associated with an increased and decreased risk of developing AD, respectively [11].

Before designing an AI model, the significant variants in the genome should be known. GWAS stands for Genome-Wide Association Studies, which is a research approach used to identify genomic variants related to a certain disease or a specific trait [12]. GWAS identifies genomic risk loci, which are sets of correlated single nucleotide polymorphisms (SNPs) that exhibit a statistically significant association with the disease or trait under investigation [13]. These studies have gained tremendous interest in finding specific genes that predispose individuals to common disease traits, most of which follow complex inheritance patterns rather than Mendelian patterns. In order to perform the genome association analysis, PLINK, a software package used for GWAS and other types of genetic analyses [14].

AI is transforming the healthcare industry due to the rising availability of unstructured and structured data and the rapid development of analytical methodologies [15]. As AI becomes more important in healthcare, there are growing worries about a lack of transparency and explainability, as well as potential bias in model projections. AI can be used to improve Alzheimer's detection and diagnosis while also minimizing overtreatment. However, merging AI with Machine Learning (ML) techniques allows for predictions and more precise decision-making. Harvard University researchers [16] have built a deep learning model that can predict Alzheimer's disease from brain scans with excellent accuracy, even in cases of the early start. The AI model was trained on a massive dataset of MRI scans and genetic data from Alzheimer's sufferers and healthy controls. The study showed that AI has the potential to improve early identification and diagnosis of Alzheimer's disease.

This work aims to contribute to AD research by integrating ML systems with GWAS to reliably detect and predict the disease. It addresses the challenge of interpretability in ML models by employing explainable ML techniques, shedding light on the features contributing to predictions. Using the Alzheimer's Disease Neuroimaging Initiative dataset and rigorous quality control methods, the study identifies strong connections between specific genetic variants and AD, highlighting the potential of GWAS in disease detection.

The remainder of this report is structured as follows: Section II represents the related works of using AI to detect Alzheimer's. AD GWAS dataset is described in Section III. Sections IV and V discuss the genome-wide association studies and quality control procedure, respectively. The concept of XAI is discussed in Section VI. The methodology for achieving the aim of the work is described in Section VII. The results of performing quality control procedures, GWAS, and ML-model evaluation are presented in Section VIII. Finally, conclusions and future work are summarized in Section IX.

II. LITERATURE REVIEW

In this section, some related works of Alzheimer's detection using ML are discussed. Abbas et al. [16] tried to identify biomarkers-related AD SNPs in order to design a deep learning-based model for AD classification. They trained convolutional neural networks (CNNs) on a GWAS dataset obtained from the AD neuroimaging initiative. Subsequently, deep transfer learning was applied to further train the CNN

as a base model on a separate AD GWAS dataset, leading to the extraction of a final set of features. These extracted features were then utilized as inputs for a Support Vector Machine (SVM) to classify AD. Extensive experiments were conducted using multiple datasets and different experimental setups. The statistical analysis revealed an accuracy of 89% for the classification of AD.

M. Menagadevi [17] proposed an Alzheimer's disease detection method that combines multiscale pooling residual autoencoder and Support Vector Machine (SVM) for analysis. It utilizes image datasets from Kaggle and ADNI, enhancing images through modified optimal curvelet thresholding and Octagon histogram equalization with black-and-white stretching. The multi-scale pooling residual autoencoder extracts relevant white matter features. For classification, Support Vector Machine (SVM), Extreme Learning Machine (ELM), and K-nearest neighbors algorithm (KNN) are employed. Notably, SVM demonstrates outstanding performance with an impressive accuracy rate of 99.77% for the Kaggle dataset and 98.21% for ADNI, highlighting its efficacy in Alzheimer's disease classification.

Abd El Hamid et al. [18] utilized Naive Bayes, K2 learning algorithms, and tree-augmented Naive Bayes. for the early detection of Alzheimer's disease. Based on genetic data from the Alzheimer's disease neuroimaging initiative phase 1 dataset, 500 SNPs were used to achieve the highest classification accuracy according to the p-value requirement, which equals 0.05. Overall accuracy for the Naive Bayes and K2 learning algorithms was 98% and 98.4%, respectively. A. Alatrany et al. [19] developed and assessed a deep learning model for Alzheimer's prediction using genetic information from 188 controls and 176 AD patients. The model achieved an area under the curve (AUC) of 0.93 and 0.09 using multilayer perceptron and convolutional neural networks, respectively. The same authors concentrated on using a layered Machine Learning (ML) based model to categorize Alzheimer's patients. The model was evaluated using all of the AD genetic data from ADNI-1 which is the first part of the neuroimaging experiment. With an overall accuracy of 93.7%, the authors claim that the stacked model performed better than conventional machine learning techniques. They indicated that stacking methods are successful in identifying Alzheimer's disease.

In order to forecast a patient's probability of developing AD, Araujo et al. [20] proposed the use of physiologically motivated SNP selection as a data point in RF. Their research indicates that SNPs can be effective as data points in RF for predicting AD risk. Importantly, the authors found that these selected SNPs, even if they are not directly linked to the disease, perform better than SNPs that are associated with AD. To identify SNPs associated with AD in a GWAS data set of 550 healthy and 861 diseased, two unique approaches were developed by N. Briones and V. Dinu [21]. In the first method, the authors utilized logistic regression to filter the data by applying a predetermined p-value threshold, producing a block of SNPs that were then used in a multi-locus study using random forest, while using biological data and logistic regression analysis to pre-select loci for input into the RF classifier in the second technique. The first method yielded 199 SNPs. These SNPs, along with other SNPs linked to AD, were used to create a predictive subgroup for AD prediction. Utilizing 10-fold cross-validation in random forest (RF) modeling, the average error rate for AD prediction was determined to be 9.8%.

GenEpi, a computational tool that uses L1-regularized regression to identify epistasis associated with phenotypes, was introduced by Chang et al [22]. For the purpose of determining both within-gene and cross-gene epistasis, GenEpi employs a two-stage modeling methodology. On the basis of 364 people's genetic information, the ML model was trained and assessed. The final model made use of 24 SNPs overall, spread across 12 genes. The model demonstrates a leave-one-out cross-validation accuracy of 0.83 and a 2-fold cross-validation accuracy of 0.83. Cooper et al. [23] evaluated the anticipated performance and efficacy of a Bayesian approach to several conventional ML techniques using a GWAS dataset of AD that consists of 312,318 SNPs of 1411 participants. The findings indicate that the Bayesian algorithm achieves comparable prediction results to conventional methods while exhibiting a reduced training time requirement.

Oriol et al. [24] used FRESA.CAD (Feature Selection Algorithms for Computer Aided Diagnosis) to predict the hereditary risk of developing AD. It is a benchmarking tool that works by building and assessing a number of ML models, such as Least Absolute Shrinkage and Selection Operator (LASSO), Bootstrap Stage-Wise Model Selection (BSWiMS), and Recursive partitioning and regression trees (RPART). The range of the AUC value was from 0.6 to 0.7. The ensemble of techniques performed best, with a receiving operation curve (ROC) score of 0.719, and was competitive with the BSWiMS, LASSO, and RPART.

Based on a selection of the 21 variants most closely associated with AD, in [25], SVM classifiers of multiple kernels were applied to the ADNI data using the correlation-based and chi-squared approaches. The findings demonstrate that an RBF kernel-based SVM-trained model has a maximum accuracy of 76.70%. To determine if the data used to describe one dataset could be successfully used to categorize a completely other patient group dataset, the authors [26] conducted two different types of experiments. In the first experiment, the authors used features chosen from the initial dataset to train a random forest classifier. The second dataset is used to assess the model results. Subsequently, the authors employed the selected SNPs locations to construct a novel random forest model using the second dataset. The feature selection process for the second dataset was conducted based on the training subset of the first dataset, focusing on relevant features. Patients from the second dataset, who had not been used in training the tested, classifier were used in both tests to assess the performance of the final classifier. In comparison to the results obtained from using a single dataset, both experiments demonstrated a slight reduction in the AUC values.

However, the AUC values remained significantly above 0.5, implying that all the models retained valuable information regarding genetic distinctions between Alzheimer's disease cases and controls.

In order to choose a subset of SNPs for GWAS, Nguyen et al. [27] suggest a new two-stage random forest technique called ts-RF. It has been discovered that the suggested technique is effective at locating educational sets of SNPs that may be connected to illnesses. Other works [28] demonstrated a novel technique for the analysis of AD using GWAS that combines both enrichment analysis and random forests to identify new genetic variants or biomarkers based on data from 527 controls and 117 cases. Romero et al.

The study in [29] proposed a deep-learning model that can find interactions between SNPs. The Deep Mixed Model is made up of two parts: the first part uses a CNN to account for confounding factors, while the second part uses an LSTM to pick genetic variations. Rosales et al. [30] compared three ML models: genetic algorithm, stepwise, and L1- regularization techniques (LASSO) for building models for predicting Alzheimer's disease based on data from 813 diseased and 1,017 healthy. LASSO models fared better than the other two methods in predicting whether the patients have AD or not.

Sherif et al. [31] devised a framework for comparing various Bayesian network methods (naive Bayes, Markov Blanket (MB), tree-augmented naive Bayes, and minimal augmented Markov blanket). For naive and tree-augmented naïve networks, a total of 435 were regarded as predictors. However, using only 11 and 13 SNPs for the minimum augmented MB and Markov blanket training and testing, respectively, demonstrated improved accuracy. AD was predicted using the model-averaged naive Bayes (MANB) approach by Wei [32]. On the basis of 1411 people's genetic data, the models were trained and tested. The model outputs were compared with the results obtained from a naive Bayes classifier. Despite having a similar training time, the model achieved a significantly higher AUC of 0.72, whereas the naive Bayes classifier yielded an AUC of 0.59.

Stokes et al. [33] evaluated the effectiveness of label propagation (LP), a multivariate graph-based method in order to efficiently rank SNPs in genome-wide data. The top-ranked SNPs were assessed based on classification accuracy and prior evidence linking them to AD. Compared to other control approaches, LP scored significantly better at categorization. Among the 25 top-ranked SNPs discovered by LP, 14 were found in one dataset and had evidence in linking them to AD.

III. AD GWAS DATASET

The inclusion criteria for participants in this study consisted of the following factors: a) self-reported European ethnicity, b) adherence to the standards set by the National Alzheimer's Coordinating Centre, and c) confirmation of late-onset Alzheimer's disease (AD) by board-certified neuropathologists for cases, while controls exhibited no neuropathology. Moreover, participants aged 65 years and above were selected for inclusion. All cases and controls underwent plaque and tangle assessment, which are distinctive structures affecting brain cells and potentially contributing to the pathophysiology of the disease. Samples with a history of stroke, Lewy bodies, or any other neurological disorder were excluded from the analysis.

The final dataset comprised 191 males and 173 females, with a total of 176 cases and 188 controls, each possessing genotyping information for 502,627 single-nucleotide polymorphisms (SNPs). Genotyping was performed on the DNA of participants using the Affymetrix GeneChip Human Mapping 500K Array Set. The onset of AD can be early or late and every type has its own genes. For instance, early-onset AD is caused by mutations in Presenilin 1 (PS1), Presenilin 2 (PS2), and Amyloid precursor protein (APP). On the other hand, late-onset AD results due to changes in the APOE gene, microtubule-associated protein tau (MAPT) gene, and tumor necrosis factor (TNF) gene. The APOE gene has 3 forms including APOE2, APOE3, and APOE4. The significant SNPs that cause the variation in the APOE gene are rs429358 and rs7412, which have the T or C allele. The AD GWAS dataset is mainly focused on late-onset AD.

IV. GENOME-WIDW ASSOCIATION ANALYSIS

SNP studies, a type of GWAS, examine the phenotypic impact of tiny genetic variants. While some approaches for GWAS analysis concentrate on phenotypic risk prediction based on the available genetic data [34] [35], others attempt to interpret these risk effects by highlighting which SNPs are influencing a particular trait [30]. In order to find SNPs connected to the phenotype under research, this study combines both of these objectives and applies a deep learning-based prediction algorithm in conjunction with statistical analysis. GWAS is a study design used to identify genetic variants associated with common human diseases and traits, such as heart disease, type 2 diabetes, and psychiatric disorders [36]. The experimental procedure of a GWAS encompasses collecting DNA and phenotypic information from a cohort of individuals, including information on disease status and demographic characteristics [37]. GWAS analyzes hundreds of thousands to millions of SNPs across the genome to identify genetic variants associated with a trait [37].

A variety of applications can make use of the GWAS results. In order to address potential confounding genetic group differences, it is common practice in epidemiological studies to incorporate trait-associated genetic variants as control variables. This helps to account for any potential biases that may arise from these genetic differences [38]. Additionally, based on a person's genetic profile, the results can be used to predict their risk for contracting physical and mental diseases. In fact, a recent study demonstrated that the prediction methods of the monogenic risk based on uncommon, highly penetrant mutations are just as effective at predicting disease risk. Genomic risk prediction methods make use of genome-wide polygenic risk scores (PRSs) for various conditions, including atrial fibrillation,

coronary artery disease, inflammatory bowel disease, type 2 diabetes, and breast cancer. These PRSs are calculated based on comprehensive genetic information obtained from across the genome and are employed to estimate an individual's predisposition or susceptibility to these specific diseases.

A. GWAS CONDUCTING

a: SELECTING STUDY POPULATIONS

To uncover replicable genome-wide significant associations, GWAS may require very high sample sizes and the desired sample size can be computed using power estimates in software programs such as CaTS14 or GPC15. When the characteristic of interest is dichotomous, different study designs can be employed. One approach involves including both cases and controls, allowing for a comparison between individuals with and without the trait. Alternatively, in cases where the trait is quantitative, quantitative measures can be collected for the entire study population to assess variations and associations with the characteristic of interest. Furthermore, there are other approaches including population-based and family-based designs. The desired size of the sample, the experimental topic, and the availability of pre-existing data or the feasibility with which new data can be obtained all influence the selection of data resources and research design for conducting the GWAS. GWAS can be carried out utilizing diverse data sources such as biobanks, disease-focused cohorts, population-based cohorts, or direct-to-consumer surveys. Recruitment tactics must be carefully evaluated for all study designs because they can cause collider bias and other types of bias in the resulting data [39]. One example of a widely used study cohort is the UK Biobank, which adopts a volunteer-based recruitment strategy. As a result, participants in the UK Biobank cohort tend to exhibit better health, higher socioeconomic status, and higher educational attainment compared to the general population.

b: GENOTYPING

Individuals are often genotyped using microarrays for common variations or next-generation sequencing technologies such as WES (whole-exome sequencing) or WGS (whole-genome sequencing) for rare variants. Due to the current expense of next-generation sequencing, microarray-based genotyping is the most often utilized approach for acquiring genotypes for GWAS. However, the choice of genotyping platform is influenced by a variety of criteria, including the objective of the GWAS; for instance, WGS, which determines nearly every genotype of a whole genome, is favored above WES and microarrays and is projected to become the method of choice in the coming years as low-cost WGS technology becomes more widely available [40].

c: DATA PROCESSING

Individual ID numbers, sex, coded family relations between individuals, covariates, phenotype information, genotype calls for all called variants, and genotyping batch information are all included in GWAS input files. Following data input, producing accurate GWAS results necessitates precise quality control procedures. Testing for associations: The biometrical model explains the genetic association theory (for further information, see Supplementary Note). In GWAS, associations are often tested using linear or logistic regression models, depending on the nature of the phenotype being investigated. Linear regression models are commonly employed for continuous phenotypes such as height, blood pressure, or body mass index. On the other hand, logistic regression models are utilized for binary phenotypes, such as determining the presence or absence of disease [40]. In order to address stratification and mitigate potential biases stemming from demographic factors, adjustments are made by including covariates like age, gender, and ancestry. However, it's important to note this may reduce the statistical strength when dealing with binary traits in selected study samples. [40].

d: ACCOUNTING FOR FALSE DISCOVERY

To avoid false positives, examining millions of connections between individual genetic variations and a phenotype of interest necessitates a strict multiple-testing threshold [40].

V. GWAS QUALITY CONTROL

Quality control (QC) [41] is a critical step in any genetic study that involves collecting, processing, and analyzing biological samples. It is the process of verifying and ensuring the quality and integrity of the data obtained from these samples. In genetic studies, quality control involves a series of steps that are performed to identify and remove low-quality or unreliable data points. This includes detecting and correcting errors in genotyping data, identifying, and removing outliers, checking for sample contamination, and ensuring that the data conforms to standard quality metrics. Common quality control procedures in genetic studies may include:

- Removing samples with a low genotyping rate or high missing data rates.
- Removing samples with unexpected genetic ancestry or relatedness.
- Removing SNPs with low call rate or high missing data rates.
- Removing SNPs with significant deviation from Hardy-Weinberg equilibrium (HWE).
- Checking for and removing duplicates or samples with low DNA quantity or quality.
- Removing SNPs with batch effects or systematic technical errors.
- Performing population stratification analysis to detect and remove outliers.

In [42], the authors detail a comprehensive description of the steps involved in data quality assessment and control during case-control association studies. The steps described involve the identification and elimination of DNA samples and markers that may introduce bias. Before statistically testing for the association, these crucial procedures are important for successfully conducting the case-control study.

They explained how to make assessments of failure rate per individual and per SNP as well as how to gauge the degree of relatedness between individuals using PLINK, a program for managing SNP data. They also go through other quality-control techniques, such as the use of SMARTPCA software to find ancestor outliers. The aim of quality control in genetic studies is to ensure that the data is reliable, consistent, and unbiased and that the results obtained from the data are valid and accurate. Proper QC procedures can improve the statistical power of the study, minimize false positives and false negatives, and increase the reproducibility of the findings. In summary, quality control is a crucial step in genetic studies to ensure that the data is of high quality and to minimize the risk of bias or errors that can affect the results of the analysis.

In this work, we performed population stratification analysis to identify and remove outliers, ensuring that our dataset accurately represented the genetic ancestry of the study population. Additionally, we checked for and removed duplicates or samples with low DNA quantity or quality, further enhancing the reliability of our dataset.

Importantly, each QC step had a direct impact on the size of the final dataset. By removing low-quality samples or SNPs, we ensured that only high-quality data points were retained for downstream analysis. While these QC procedures resulted in a reduction in dataset size, they were essential for maintaining data integrity and minimizing the risk of bias or errors that could affect the validity of our findings.

VI. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable artificial intelligence (XAI) [43] is a set of approaches and strategies for explaining the consequences of ML model building in a way that humans can understand. The question is why explainable Machine learning is needed and why it is so important [44]. The response to "What is the accuracy" could be useless without the addition of "why we get this accuracy", therefore this is the interpretation of how the model produces the results. Three main applications of machine learning models that often involve prediction and require interpretability are model debugging, model validation, and knowledge discovery.

XAI has 2 main approaches including the intrinsic approach and the model agnostic one. In the intrinsic technique the internal parameters of the model are utilized to get explanations. On the other hand, the model agnostic approach is mainly for black box models and the internal parameters are unkown. There are various types of explanations such as intrinsic or post hoc, model-specific or model-agnostic, and global or local explanations. Model-agnostic methods [45] are powerful techniques for generating explanations without relying on the internal workings of machine learning (ML) models, which can often be opaque or difficult to interpret. One key advantage of these methods is their ability to be applied to any ML model, irrespective of its architecture or complexity. This versatility allows researchers and practitioners to employ model-agnostic methods across a wide range of ML models, enhancing transparency and interpretability in the decision-making process. One example of a model-agnostic approach [46] is feature importance analysis.

Feature importance refers to the process of identifying the most significant features or variables that contribute to the performance of a model. There are many different methods for feature importance analysis, such as permutation feature importance, mean decrease impurity, and SHAP values and these techniques can be applied to any model regardless of the specific algorithm used. Permutation importance allows the identification of the most important features [44]. It is based on shuffling the values of a feature and repeating the prediction while monitoring the error. If the error worsens, this means that this feature is important and highly impacts the prediction. Hence, the more important the specific feature, the more the predictions will worsen because of the shuffling. Hence, this method ranks the SNPs in our data from the most important one to the least important.

A partial dependence plot (PDP) offers insights into how specific features influence predictions. It is a graphical representation illustrating the relationship between one or more input variables and the output target. By examining a PDP, we can discern how alterations in predictions are influenced by the most significant features. From the PDP plots, we can know if there is a linear relationship between the predicted AD and any one of the SNP genotype values. In addition, interact PDP helps in this framework by investigating the interaction between two SNPs and their effect on the model prediction.

Furthermore, Two widely used methods for model interpretability and explainability in machine learning are SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) [47]. The way LIME works is to first pick a sample to interpret. The objective is to repeatedly test the model to understand how it generates the prediction for the selected example [47]. LIME produces local explanations by locally approximating the model using a simpler model (such as a linear model) and manipulating the input data to observe how the output changes. This method can be applied to any model because it is model-agnostic. The global behavior of the model or interaction between characteristics is not taken into account by LIME, which only offers local explanations, unlike SHAP which provides global explanations. SHAP explanations are a popular feature-attribution technique for explainable AI. They quantify the impact of specific features on the forecast of a machine-learning model using ideas from game theory [48].

By incorporating SHAP and LIME methods into our framework, we aim to provide both global and local insights into the predictions of our ML model for Alzheimer's disease detection. These explanations can enhance the interpretability of the model's decisions, foster trust in its predictions, and facilitate further research into the underlying genetic factors contributing to the disease. Fig 1 depicts the complete process of building the ML model starting from training the model followed by the evaluation of the model and ending with the interpretation of the results This framework serves

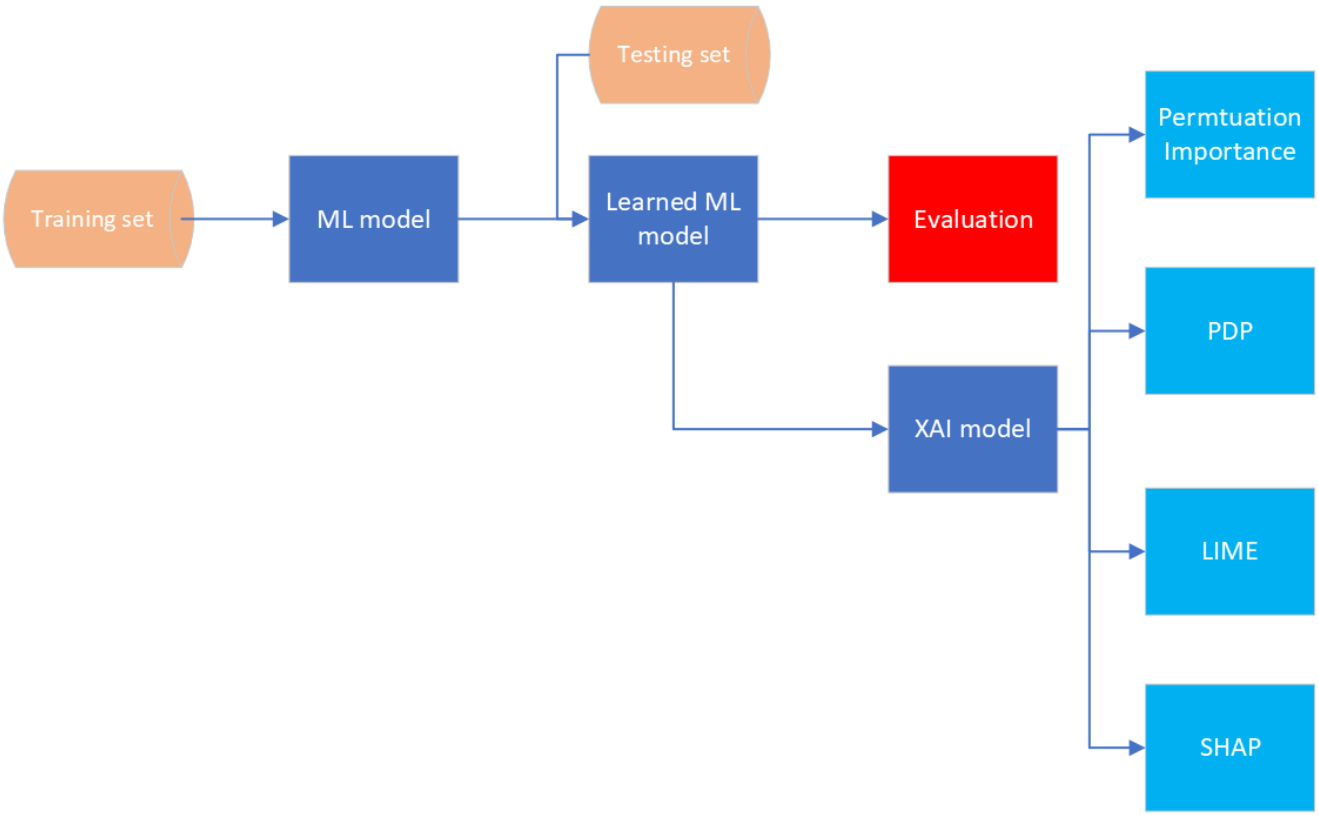

FIGURE 1: Simple workflow for developing an AI model.

as a valuable guide for researchers and practitioners in the field of ML to build robust and accurate models.

VII. METHODOLOGY

The methodology for detecting Alzheimer's disease using ML based on genetic data can be summarized in the following steps:

1) **Data Collection:** The first step is to collect genetic data of individuals that includes their DNA sequencing, Single Nucleotide Polymorphism (SNP) data, and clinical information of Alzheimer's disease.
2) **Quality Control:** The raw genetic data undergoes a series of QC checks to ensure that the data is reliable and accurate. QC steps may include filtering out SNPs with low call rates, removing individuals with high rates of missing genotype data, checking for population stratification, performing identity-by-descent (IBD) analysis to identify cryptic relatedness, and calculating the Hardy-Weinberg Equilibrium (HWE).
3) **Genome-wide Association Study (GWAS):** GWAS analysis is performed to identify genetic variants that are associated with Alzheimer's disease. GWAS analysis involves testing millions of SNPs across the genome for association with the disease. SNPs that reach genome-wide significance are then considered for further analysis.
4) **Data Preprocessing:** Preprocessing the data includes cleaning, normalization, and transformation of genetic data. The QC data needs to be preprocessed and formatted to remove errors, and inconsistencies and reduce noise. This step is critical as the quality of the data directly impacts the performance of the ML model.
5) **Feature Selection:** The next step is to select the relevant features from the genetic data that can help in the detection of Alzheimer's disease. Feature selection can be performed using statistical methods or domain knowledge. This process helps in reducing the dimensionality of the data, which improves the efficiency and accuracy of the ML algorithm. In this work, the Top 25 SNPs are selected to train our ML model. By focusing on these SNPs, we prioritize the inclusion of features that have shown the strongest evidence of association with AD risk. Then, we ended up with 364 patients with 25 SNPs as features. Out of the 364, 190 are control (without AD) and 174 are cases (with AD).
6) **ML Model Selection:** The next step is to select the appropriate ML algorithm that can effectively classify individuals with or without Alzheimer's disease based on genetic data. Commonly used ML algorithms for disease detection include SVM, Random Forest, and Neural Networks.

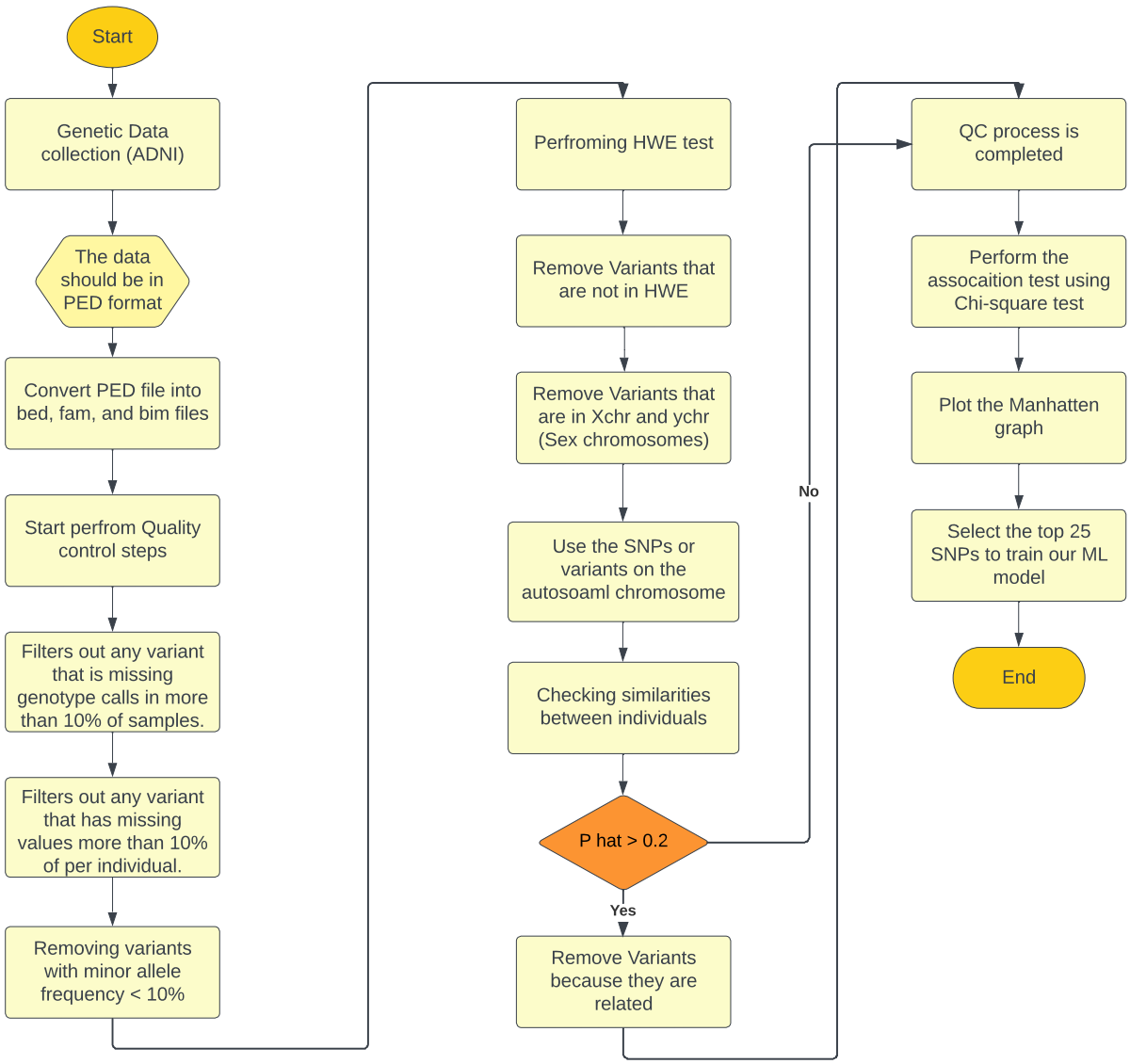

FIGURE 2: Complete Process of Building Alzheimer predictive AI Based-Model.

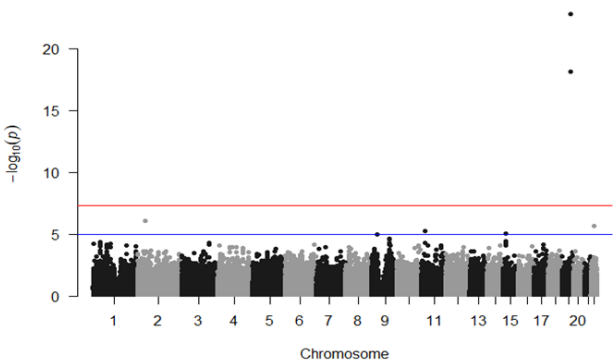

FIGURE 3: Manhatten plot.

7) **Model Training:** In this step, the ML algorithm is trained using the preprocessed data with selected features. The training process involves feeding the data to the algorithm, and the algorithm learns the patterns and relationships between the features and the output (Alzheimer's or non-Alzheimer's).
8) **Model Evaluation:** The performance of the trained ML model is evaluated using various performance metrics such as accuracy, precision, recall, and F1-score. This step helps in determining the effectiveness of the model in detecting Alzheimer's disease based on genetic data.
9) **Model explanation:** The model is interpreted using model-agnostic methods to better understand the model behavior.
10) **Model Deployment:** Finally, the trained ML model is deployed for use in real-world scenarios. The model is

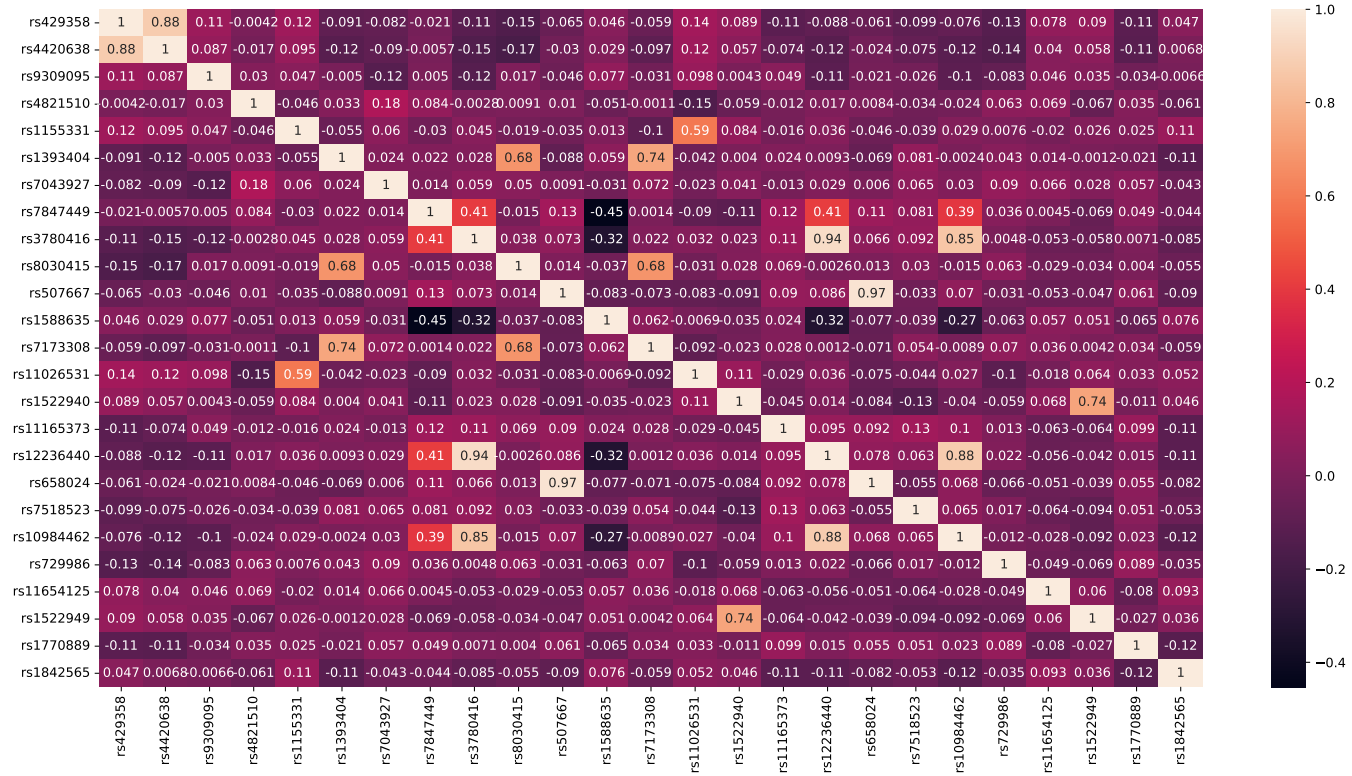

FIGURE 4: Correlation matrix for the dataset.

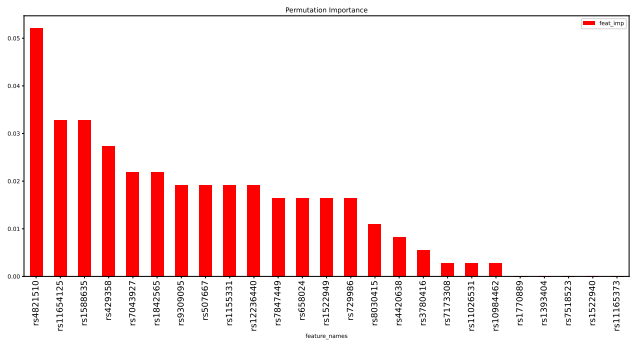

FIGURE 5: Permutation importance plot.

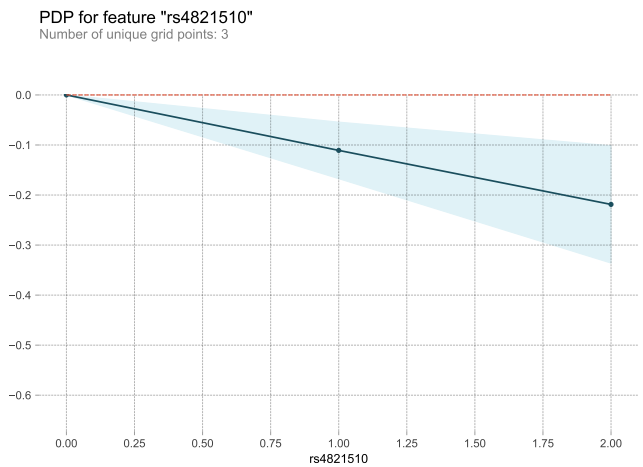

FIGURE 6: PDP for rs4821510 SNP.

aimed to be integrated with clinical diagnostic tools to provide early diagnosis of Alzheimer's disease and aid in personalized treatment plans.

Predicting Alzheimer's disease using ML based on genetic data involves collecting and preprocessing the data, performing QC checks, selecting relevant features, performing GWAS analysis, choosing an appropriate ML algorithm, training the model, evaluating and interpreting its performance, and deploying the model for use in clinical settings. Fig 2 shows the flow chart for predicting AD based on genetic data.

VIII. SIMULATION RESULTS

This section presents the results obtained from the three processes including quality control, genome-wide association studies, and machine learning. The quality control and GWAS procedures were performed using R software while Python software was utilized to train and evaluate the ML model.

TABLE 1: Classification Results of AD Recognition.

Models	Hyper - Parameters	Precision	F1-score	Computational time
SVM	Kernal = Linear C = 1	**89.00%**	**0.89**	0.016 s
RF	No. of estimators = 100 criterion = Gini	87.5%	0.88	0.103 s
MLP	No. layers = 4 Activation = ReLU solver='lbfgs' Learning rate = 1e-5,	85%	0.85	0.0737
KNN	No. neighbours = 5 metric = Minkowski	97%	0.775	0.0009 s
LightGBM	No. of threads = 0 objective = binary	88 %	0.88	0.042 s
Adaboost	No. of estimators = 100 algorithm = SAMME	85%	0.85	0.1013 s

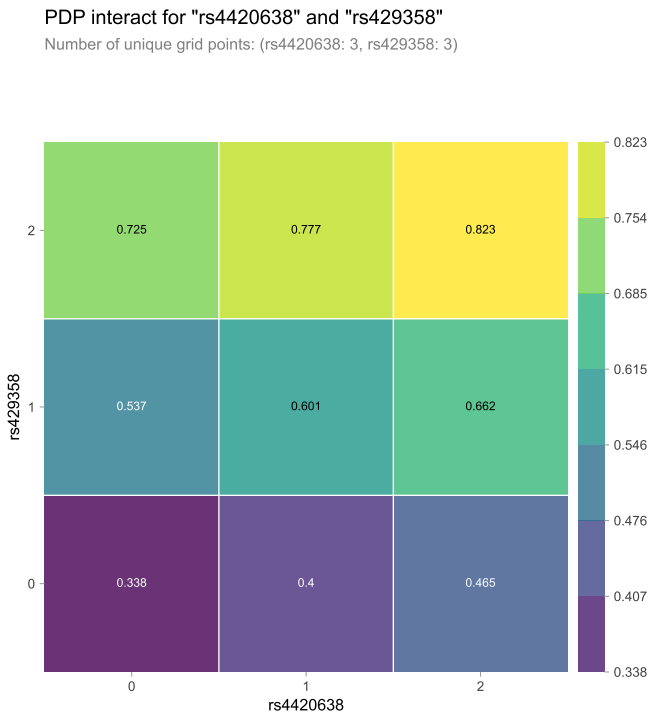

FIGURE 7: Interaction PDP plot for rs429358 and rs4420638 SNPs.

A. QC AND GWAS RESULTS

The results of the first step in the QC indicated that 4804 variants were removed due to missing genotype data and 375353 variants for 364 people passed the filter. Then, 33338 variants were removed due to the HWE test as well as 84025 variants because they have a minor allele frequency which is less than 10%. Hence, 257990 SNPs for 364 people remained. The step of checking the sex discrepancy revealed that there are 5392 SNPs on the X chromosome and 0 SNP on the Y chromosome which ensures that the data has both men and women. On top of that, the other variants or SNPs which are about 252490 are on autosomal chromosomes. The final step in the QC is to remove the related variants and the results showed that there are no related variants. Therefore, the number of SNPs that will come under the association test is 252490 SNPs. Next, an association test was conducted to assess each SNP and assign a p-value. In this work, a basic allele-based chi-squared association test was utilized to find the association between the SNPs and Alzheimer's disease. In order to figure out the most significant SNPs, the Manhattan plot was graphed to recognize the important SNPs. Fig 3 shows the Manhattan plot and it reveals that two SNPs pass the red line which means they passed the GWAS significant threshold hence, these two SNPs, rs429358 and rs4420638, are associated with AD. GWAS results show that there is a linkage between these two SNPs, indicating a higher likelihood of being inherited together.

B. ML MODEL PERFORMAMCE

The top 25 SNPs were selected for the 346 people to train an ML model to predict AD. First, a correlation matrix is performed to find the correlation between the SNPs. From Fig 4, rs658024 is highly correlated with rs507667, and rs780416 is highly correlated with rs12236440 as well. In our study, the training and test data ratio was 80:20. Despite the small dataset size, we believe this split ratio allowed us to effectively train and evaluate our models while ensuring sufficient data for testing. Furthermore, we conducted multiple experiments to ensure the robustness of our results and verified the consistency of our findings across different runs, and the results shown are not overfitted. Hence, Various ML algorithms were used to train the model including SVM, Random Forest, KNN, and XG-boost. The best-performing model was SVM which achieved 89% accuracy. Table 1 summarizes the performance, hyperparameters, and computational time for all the used models.

The values for precision and F1-score suggest that the

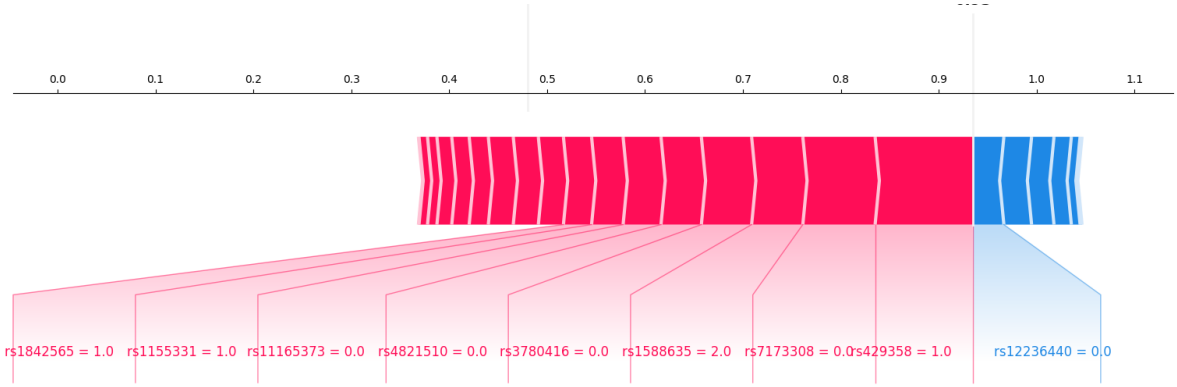

FIGURE 8: SHAP plot for the healthy (control) class.

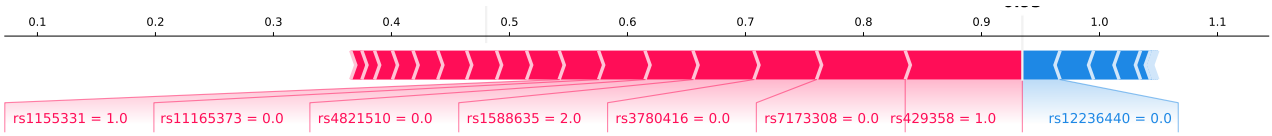

FIGURE 9: SHAP plot for the cases class

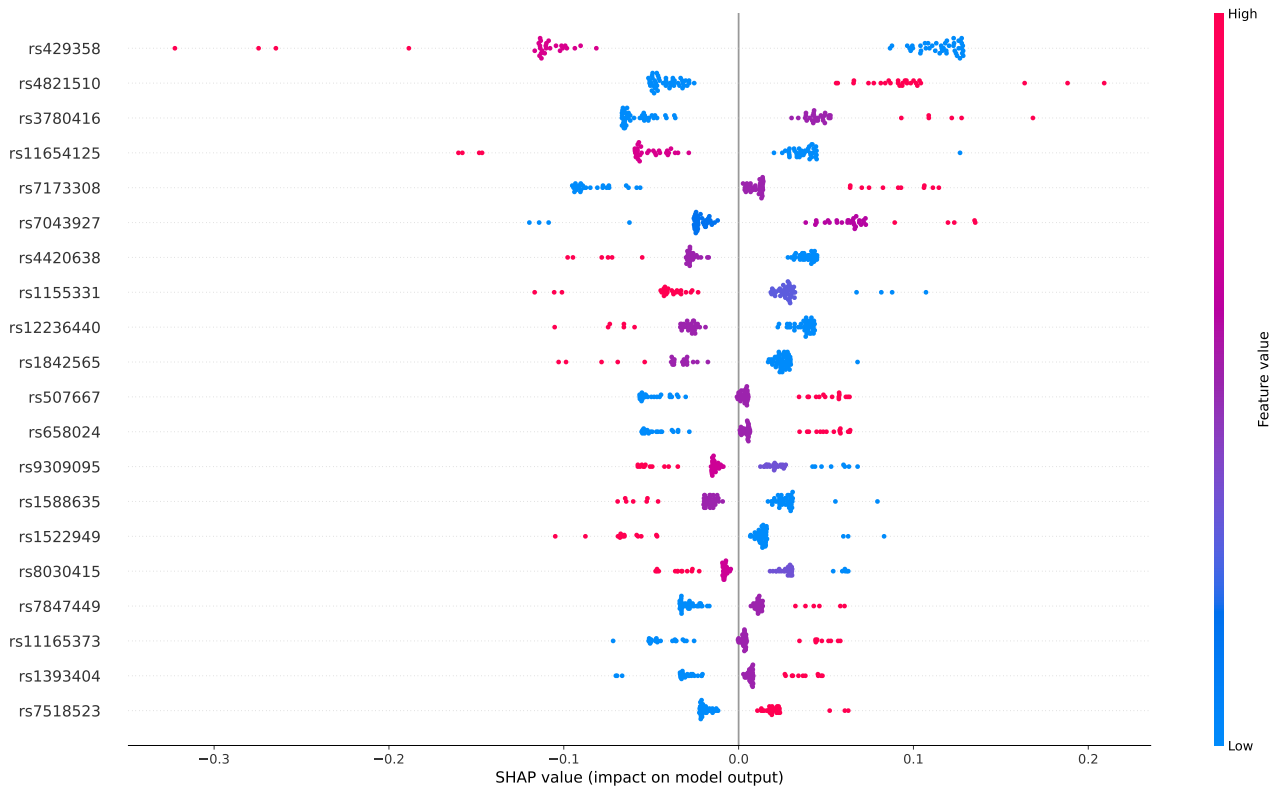

FIGURE 10: SHAP summary plot for the class 1 (Control)

SVM classification model is performing reasonably well for both classes, with a relatively balanced performance between precision and recall, as confirmed by the similar F1-score for both classes. It is important to know that the performance of a machine learning algorithm is affected by a variety of factors, such as data quality, feature selection, hyperparameter

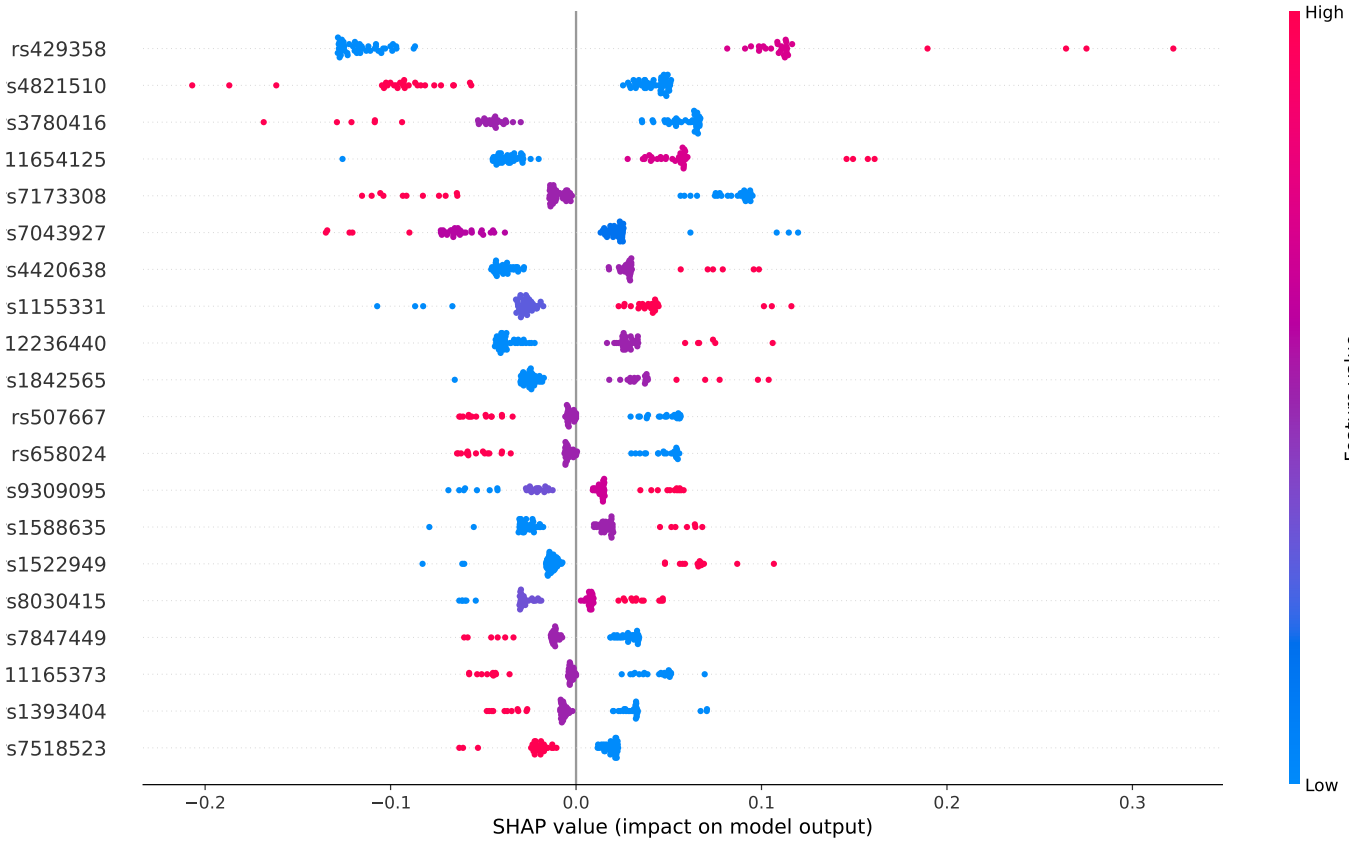

FIGURE 11: SHAP summary plot for the class 2 (Case)

tuning, and the evaluation measure utilized.

C. XAI RESULTS

To complete the cycle of machine learning, after getting the ML results, explaining and interpreting the results should be the final step. Model-agnostic methods including the permutation importance method, LIME, and SHAP have been used to provide explanations for the model outcomes. The Permutation Importance Method offers a coarse measure of feature importance, making it relatively easy to interpret but lacking the detail provided by LIME and SHAP. LIME focuses on local interpretations, making it suitable for understanding individual predictions, while SHAP offers both local and global explanations, providing a more comprehensive view of model behavior. However, SHAP's insights, grounded in game theory, may require additional expertise to interpret. In terms of computational cost, the Permutation Importance Method is computationally lightweight, while LIME and SHAP may require more resources, especially for large datasets or complex models. Thus, the choice of method depends on the specific goals of the analysis and the desired level of detail and interpretability. Combining multiple methods can offer complementary insights and enhance the understanding of model behavior.

Fig 5 shows that according to the shuffling of the values of the features (genotype values) of the SNPS, rs4821510 is the most important SNP for the SVM classifier in detecting AD. In order to demonstrate the impact of the rs4821510 SNP on the model behavior, a PDP is presented in Fig 6 which shows that when the genotype varies from 0 to 2, the likelihood of the model predicting the diseased class decreases.

As mentioned earlier, the analysis indicates that the SNPs rs429358 and rs4420638 are likely inherited together, hence a PDP interaction plot can show the results of such interaction. Fig 7 depicts that when the genotype value of rs4420638 and rs429358 is 2, representing the TT allele, the probability of the model predicting the diseased class increases. Furthermore, the plot demonstrates that even when rs429358 is TT, the ML model tends to predict AD cases regardless of the genotype value of the rs4420638 SNP.

SHAP can produce local explanations for the ML results. Therefore, we selected a single instance from the dataset with a health class as a target as shown in Fig 8. It is clear that when rs429358 SNP has a 0 genotyping value, the model tends to predict the healthy class.

In Fig 9, the SHAP plot depicts that when rs429358 SNP has a 1 genotype value, the model is likely to predict the diseased class. SHAP produces global explanations for the ML results using kernel explainer. The SHAP plot summary in Fig 10 shows that rs4821510 SNP affects positively the model detection of class 1 which is the health cases, unlike the rs429358 SNP which affects negatively. This indicates

that if the rs4821510 genetic variant is present, the model will likely predict health cases. On the other hand, the presence of rs429358 SNP relates to the diseased cases which is shown in Fig 11.

As a result of this, we can identify genetic markers linked to Alzheimer's disease and utilize them to train the ML model by applying GWAS and quality control. Considering the results and interpretations included in this study, this approach is shown promising for enhancing Alzheimer's identification and may help in designing earlier therapies to enhance patient outcomes.

While Permutation Importance Method provides valuable insights into feature importance, multiple iterations may increase runtime, presenting a limitation, especially for large datasets [49]. Additionally, LIME's applicability is restricted to supervised Machine Learning and Deep Learning models, limiting its versatility. On the other hand, global SHAP methods like KernelSHAP can be computationally slow due to the need to compute Shapley values for numerous instances, posing a challenge, particularly for complex models or extensive datasets [50]. These limitations underscore the importance of considering computational efficiency and model compatibility when selecting interpretation techniques, ensuring that the chosen method aligns with the specific requirements and constraints of the analysis.

IX. CONCLUSION

AI and GWAS can be considered effective combination for the prediction of AD. This paper presented the results of a study that used quality control measures to improve the ADNI dataset and GWAS techniques to identify genetic variants associated with Alzheimer's disease. The findings of this study suggest that these methods can be used to identify new genetic targets for the development of treatments for Alzheimer's disease. The application of ML algorithms on the dataset provided a method for identifying patients with Alzheimer's disease with high accuracy.An ML model was trained to classify patients with Alzheimer's disease and healthy controls based on their genetic data. The best-performing model was SVM, achieving 89% accuracy. The results of applying XAI showed that rs4821510 SNP and rs429358 SNP play an important role in the detection of AD. A partial dependence plot demonstrates that as the genotype ranges from 0 to 2, the probability of the model predicting the diseased class diminishes. Additionally, the interaction PDP plot indicates that when rs429358 is TT, the ML model tends to predict AD cases irrespective of the genotype value of the rs4420638 SNP. Moreover, the SHAP method reveals that the presence of the rs4821510 genetic variant strongly suggests that the model will predict healthy cases, while the presence of the rs429358 SNP is associated with diseased cases. These findings suggest that the combination of quality control, GWAS, and ML techniques can be considered as powerful approach for detecting and predicting Alzheimer's disease, providing a potential avenue for earlier diagnosis and treatment. However, future research should focus on validating the findings across diverse populations and integrating additional data sources to enhance predictive accuracy. Prospective clinical studies are needed to assess the real-world performance and feasibility of implementing the model in clinical practice.

REFERENCES

[1] K. Shanmugavadivel, V. Sathishkumar, J. Cho, and M. Subramanian, "Advancements in computer-assisted diagnosis of alzheimer's disease: A comprehensive survey of neuroimaging methods and ai techniques for early detection," Ageing Research Reviews, p. 102072, 2023.
[2] S. L. Leal and M. A. Yassa, "Perturbations of neural circuitry in aging, mild cognitive impairment, and alzheimer's disease," Ageing research reviews, vol. 12, no. 3, pp. 823–831, 2013.
[3] P. Scheltens, K. Blennow, M. M. Breteler, B. De Strooper, G. B. Frisoni, S. Salloway, and W. M. Van der Flier, "Alzheimer's disease," The Lancet, vol. 388, no. 10043, pp. 505–517, 2016.
[4] A. Association, "2019 alzheimer's disease facts and figures," Alzheimer's & dementia, vol. 15, no. 3, pp. 321–387, 2019.
[5] "https://www.thenationalnews.com/health/2022/09/14/more-dementia-care-urged-for-middle-east-as-case-numbers-forecast-to-soar/, accessed: 17th september, 2023."
[6] L. Backès et al., "Master thesis: Diagnosis of neurodegenerative diseases with deep learning: The case of alzheimer's disease," 2022.
[7] "Rodriguez et al., harvard medical school, https://hms.harvard.edu/news/ai-alzheimers-disease, (accessed may. 10, 2023)."
[8] S. Ansari, K. A. Alnajjar, S. Mahmoud, R. Alabdan, H. Alzaabi, M. Alkaabi, and A. Hussain, "Blind source separation based on genetic algorithm-optimized multiuser kurtosis," in 2023 46th International Conference on Telecommunications and Signal Processing (TSP), 2023, pp. 164–171.
[9] C. Fabrizio, A. Termine, C. Caltagirone, and G. Sancesario, "Artificial intelligence for alzheimer's disease: promise or challenge?" Diagnostics, vol. 11, no. 8, p. 1473, 2021.
[10] R. Mishra and B. Li, "The application of artificial intelligence in the genetic study of alzheimer's disease," Aging and disease, vol. 11, no. 6, p. 1567, 2020.
[11] Y.-Y. Huang, Y.-X. Yang, H.-F. Wang, X.-N. Shen, L. Tan, J.-T. Yu, and A. D. N. Initiative, "Genome-wide association study identifies apoe locus influencing plasma p-tau181 levels," Journal of Human Genetics, vol. 67, no. 8, pp. 459–463, 2022.
[12] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine et al., "Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," Alzheimer's & dementia, vol. 7, no. 3, pp. 280–292, 2011.
[13] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma, "Genome-wide association studies," Nature Reviews Methods Primers, vol. 1, no. 1, p. 59, 2021.
[14] "https://www.cog-genomics.org/plink/2.0/input."
[15] S. Ansari and K. A. Alnajjar, "Multi-hop genetic-algorithm-optimized routing technique in diffusion-based molecular communication," IEEE Access, vol. 11, pp. 22 689–22 704, 2023.
[16] A. Alatrany, A. Hussain, J. Mustafina, and D. Al-Jumeily, "A novel hybrid machine learning approach using deep learning for the prediction of alzheimer disease using genome data," in Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part III 17. Springer, 2021, pp. 253–266.
[17] M. Menagadevi, S. Mangai, N. Madian, and D. Thiyagarajan, "Automated prediction system for alzheimer detection based on deep residual autoencoder and support vector machine," Optik, vol. 272, p. 170212, 2023.
[18] M. M. Abd El Hamid, M. S. Mabrouk, and Y. M. Omar, "Developing an early predictive system for identifying genetic biomarkers associated to alzheimer's disease using machine learning techniques," Biomedical Engineering: Applications, Basis and Communications, vol. 31, no. 05, p. 1950040, 2019.
[19] A. S. Alatrany, A. Hussain, M. Jamila, and D. Al-Jumeiy, "Stacked machine learning model for predicting alzheimer's disease based on ge-

netic data," in 2021 14th International Conference on Developments in eSystems Engineering (DeSE). IEEE, 2021, pp. 594–598.
[20] G. S. Araújo, M. R. Souza, J. R. M. Oliveira, and I. G. Costa, "Random forest and gene networks for association of snps to alzheimer's disease," in Advances in Bioinformatics and Computational Biology: 8th Brazilian Symposium on Bioinformatics, BSB 2013, Recife, Brazil, November 3-7, 2013, Proceedings 8. Springer, 2013, pp. 104–115.
[21] N. Briones and V. Dinu, "Data mining of high density genomic variant data for prediction of alzheimer's disease risk," BMC medical genetics, vol. 13, no. 1, pp. 1–12, 2012.
[22] Y.-C. Chang, J.-T. Wu, M.-Y. Hong, Y.-A. Tung, P.-H. Hsieh, S. W. Yee, K. M. Giacomini, Y.-J. Oyang, and C.-Y. Chen, "Genepi: gene-based epistasis discovery using machine learning," BMC bioinformatics, vol. 21, pp. 1–13, 2020.
[23] G. F. Cooper, P. Hennings-Yeomans, S. Visweswaran, and M. Barmada, "An efficient bayesian method for predicting clinical outcomes from genome-wide data," in AMIA Annual Symposium Proceedings, vol. 2010. American Medical Informatics Association, 2010, p. 127.
[24] J. D. V. Oriol, E. E. Vallejo, K. Estrada, J. G. T. Peña, A. D. N. Initiative et al., "Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data," BMC bioinformatics, vol. 20, 2019.
[25] M. M. Abd El Hamid, Y. M. Omar, and M. 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.
[26] M. Osipowicz, B. Wilczynski, M. A. Machnicka, and A. D. N. Initiative, "Careful feature selection is key in classification of alzheimer's disease patients based on whole-genome sequencing data," NAR Genomics and Bioinformatics, vol. 3, no. 3, p. lqab069, 2021.
[27] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, and M. J. Li, "Genome-wide association data classification and snps selection using two-stage quality-based random forests," in BMC genomics, vol. 16. Springer, 2015, pp. 1–11.
[28] Z. Liang, H. Qiong, L. Ao, and W. MingHui, "A genome-wide association study of alzheimer's disease using random forests and enrichment analysis," SCIENCE CHINA-LIFE SCIENCES, vol. 55, no. 7, pp. 618–625, 2012.
[29] H. Wang, T. Yue, J. Yang, W. Wu, and E. P. Xing, "Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies," BMC bioinformatics, vol. 20, pp. 1–11, 2019.
[30] B.-L. Romero-Rosales, J.-G. Tamez-Pena, H. Nicolini, M.-G. Moreno-Treviño, and V. Trevino, "Improving predictive models for alzheimer's disease using gwas data by incorporating misclassified samples modeling," PloS one, vol. 15, no. 4, p. e0232103, 2020.
[31] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering alzheimer genetic biomarkers using bayesian networks," Advances in bioinformatics, vol. 2015, 2015.
[32] 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, 2011.
[33] M. E. Stokes, M. M. Barmada, M. I. Kamboh, and S. Visweswaran, "The application of network label propagation to rank biomarkers in genome-wide alzheimer's data," BMC genomics, vol. 15, pp. 1–13, 2014.
[34] S. Choi, S. Bae, and T. Park, "Risk prediction using genome-wide association studies on type 2 diabetes," Genomics & informatics, vol. 14, no. 4, p. 138, 2016.
[35] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. M. Visscher et al., "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," The American Journal of Human Genetics, vol. 92, no. 6, pp. 1008–1012, 2013.
[36] U. Roshan, S. Chikkagoudar, Z. Wei, K. Wang, and H. Hakonarson, "Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest," Nucleic acids research, vol. 39, no. 9, pp. e62–e62, 2011.
[37] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," The American Journal of Human Genetics, vol. 101, no. 1, pp. 5–22, 2017.
[38] V. Q. Truong, J. A. Woerner, T. A. Cherlin, Y. Bradford, A. M. Lucas, C. C. Okeh, M. K. Shivakumar, D. H. Hui, R. Kumar, M. Pividori et al., "Quality control procedures for genome-wide association studies," Current Protocols, vol. 2, no. 11, p. e603, 2022.
[39] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," Nature protocols, vol. 5, no. 9, pp. 1564–1573, 2010.
[40] F. B. Davoud Torkamaneh, "Genome-wide association studies (gwas)." Springer, 2022, vol. 2481.
[41] M. V. Holmes, M. Ala-Korpela, and G. D. Smith, "Mendelian randomization in cardiometabolic disease: challenges in evaluating causality," Nature Reviews Cardiology, vol. 14, no. 10, pp. 577–590, 2017.
[42] M. Pirinen, P. Donnelly, and C. C. Spencer, "Including known covariates can reduce power to detect genetic effects in case-control studies," Nature genetics, vol. 44, no. 8, pp. 848–851, 2012.
[43] T. Khater, A. Hussain, R. Bendardaf, I. M. Talaat, H. Tawfik, S. Ansari, and S. Mahmoud, "An explainable artificial intelligence model for the classification of breast cancer," IEEE Access, pp. 1–1, 2023.
[44] L. Gianfagna and A. Di Cecco, Explainable AI with python. Springer, 2021.
[45] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," arXiv preprint arXiv:1606.05386, 2016.
[46] ——, "" why should i trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[47] A. Thampi, Interpretable AI: Building explainable machine learning systems. Simon and Schuster, 2022.
[48] G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the tractability of shap explanations," Journal of Artificial Intelligence Research, vol. 74, pp. 851–886, 2022.
[49] J. Petch, S. Di, and W. Nelson, "Opening the black box: the promise and limitations of explainable machine learning in cardiology," Canadian Journal of Cardiology, vol. 38, no. 2, pp. 204–213, 2022.
[50] Y. Lu, X. Fan, Y. Zhang, Y. Wang, and X. Jiang, "Machine learning models using shapley additive explanation for fire risk assessment mode and effects analysis of stadiums," Sensors, vol. 23, no. 4, p. 2151, 2023.

TAREK KHATER was born in Texas, United States of America. He received a B.S. degree in biomedical engineering from Minia University, Egypt, in 2021. He is a Research and Teaching Assistant at the University of Sharjah. Currently, he is completing his MSc in biomedical engineering at the University of Sharjah with a thesis titled explainable artificial intelligence in the classification of breast cancer in young women. He has a number of publications in the research area of machine learning, explainable artificial intelligence, and their applications in biomedical engineering.

SAM ANSARI received the Ph.D. degree in Electrical and Computer Engineering from the University of Sharjah, Sharjah, United Arab Emirates. He holds a B.Sc. degree in Telecommunication Engineering from Canadian University Dubai, Dubai, United Arab Emirates, and an M.Sc. degree in Electrical and Computer Engineering from Abu Dhabi University, Abu Dhabi, United Arab Emirates. His research interests include wireless communications, signal and image processing, molecular communication, artificial intelligence, machine learning, and renewable energy.

ABBAS SAAD ALATRANY received the BSc degree from Al-Mustansiriya University, Iraq, in 2017, and the PG diploma degree from the University of Leicester, U.K. in 2020. He is currently working towards the PhD degree in computer science with Liverpool John Moores University, U.K. His research interests include computational biology and bioinformatics.

HAYA ALASKAR is an assistant professor in the College of Computer Science and Engineering at Prince Sattam Bin Abdulaziz University, Saudi Arabia. She gained her Ph.D. in computer science from Liverpool John Moores University in 2014. She also gained an M.Sc. in applied artificial intelligence from the University of Exeter in 2009. Her research interests include artificial intelligence applications and data science. She has a number of publications concentrated on using machine learning in various medical data such as signals and images.

SOLIMAN MAHMOUD (Senior Member, IEEE) was born in Cairo, Egypt, in 1971. He received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from the Electronics and Communications Department, Cairo University, Egypt, in 1994, 1996, and 1999, respectively. He is currently a Professor in the Electrical Engineering Department at the University of Sharjah, Sharjah, United Arab Emirates. He, a distinguished academic leader in the field of Electrical Engineering, will be on leave from the University of Sharjah and will assume the position of Vice Chancellor for Academic Affairs at the University of Khorfakkan, Khorfakkan, United Arab Emirates, effective from the first of September 2023. He supervised three Ph.D. students, 15 M.Sc. students, and more than 50 senior design projects. He is actively engaged in scholarly research work and has authored or co-authored more than 170 journals and conference publications since joining academia in 1996. He received "The German-Egyptian Research Fund" Grant. The Grant has been used to finance the project "Design of CMOS Field Programmable Analog Array and Its Applications." The project has been carried out in collaboration with the Ulm Microelectronics Institute, Ulm University, Germany. His articles received more than 2100 citations, and his Google Scholar H-index is 25 (His h-index from Scopus is 22). He published six refereed research books. His research interests include mixed analog/digital integrated electronic circuit (IC) design, including mixed mode (voltage/current) analog circuits IC design, mixed (analog/digital) programmable CMOS electronics systems, biomedical circuits, field programmable analog arrays (FPAAs), and multi-standard wireless receiver design. In 2005, he received the Science Prize in Advanced Engineering Technology from the Academy of Scientific Research and Technology, Higher Ministry of Education, Cairo. He received the Distinguished Research Award from the University of Sharjah, from 2011 to 2012 and from 2014 to 2015.

AYAD TURKY received the Ph.D. degree in computer science and IT from the School of Science, RMIT University, Melbourne, VIC, Australia. He is currently an Assistant Professor of computer science with the College of Computing and Informatics, University of Sharjah. He has published more than 30 papers in international journals and peer-reviewed conferences. His current research interests include the design and development of hyper-heuristic frameworks, deep learning, machine learning, evolutionary computation, and hybrid algorithms, with a specific interest in big data optimization problems, cloud computing, dynamic optimization, and data-mining problems.

HISSAM TAWFIK (Senior Member, IEEE) is a Professor of Artificial Intelligence. He holds a Ph.D. in Computer Engineering from the University of Manchester (UK) and has a long research track record of refereed publications in reputable international journals and conference proceedings in Applied Artificial Intelligence, Biologically Inspired Systems, Big Data and Machine Learning. Hissam is currently a professor at the College of Engineering at the University of Sharjah University (UAE). Prior to that, he worked for over 20 years for various universities in the UK. Hissam served as an editor of the International Journal of Future Generation Computer Systems (Elsevier), the International Journal of Neural Computing and Applications (Springer), and International Journal of Reliable Intelligent Environments (Springer). He is a chair of the International Conference Series on Developments in eSystems Engineering (DESE) and a guest editor the journal special issue on "Robotics, Sensors and Industry 4.0" for Sensors MPDI.

EQAB ALMAJALI (Member, IEEE) received the B.Sc. degree (Hons.) from Mu'tah University, Jordan, and the M.A.Sc. and Ph.D. degrees (Hons.) in electrical engineering from the University of Ottawa, Ottawa, ON, Canada, in 2010 and 2014, respectively.,He has been an Assistant Professor with the Electrical Engineering Department, University of Sharjah, since August 2017. Prior to that, he was a Postdoctoral Fellow with the Electronics Department, Carleton University, Canada. He is the author of more than 50 technical publications and two book chapters. His current research interests include frequency selective surfaces, millimeter-wave MIMO antennas, reconfigurable antennas, RF passive and active sensors, THz antennas, and wireless power transfer. He received the prestigious Canadian National Science and Engineering Research Council (NSERC) Postdoctoral Fellowship for his research excellence, in 2014, and also received the NSERC-PGS Scholarship during his doctoral studies, in 2012.

ABIR HUSSAIN (Senior Member, IEEE) received the Ph.D. degree from the University of Manchester (UMIST), U.K., in 2000, with a thesis "Polynomial Neural Networks for Image and Signal Processing." She is currently a Professor of image and signal processing with the Electrical Engineering Department at the University of Sharjah, Sharjah, United Arab Emirates. She is also a Visiting Professor of machine learning with Liverpool John Moores University, U.K. She was involved with higher order and recurrent neural networks and their applications to e-health and medical image compression techniques. She has developed with her research students a number of recurrent neural network architectures. She is a Ph.D. supervisor and an external examiner for research degrees, including Ph.D. and M.Phil. students. She is one of the initiators and chairs of the development of e-Systems Engineering (DeSE) series. Her research interests include neural networks, signal prediction, telecommunication fraud detection, and image compression. She has published numerous refereed research papers in conferences and journals in the research areas.