INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN: 2289-8522 e-ISSN: 2180-0650
VOL. 7, ISSUE 2, 26 – 43
DOI: https://doi.org/10.15282/ijsecs.7.2.2021.4.0087

ORIGINAL ARTICLE

Machine learning and deep learning-based approaches on various biomarkers for Alzheimer's disease early detection: A review

Ghada M. Fadhl Alqubati[1,*] and Ghaleb H. Algaphari[1]

[1]Computer and Information Technology Faculty, Sana'a University, Yemen.

ABSTRACT – Alzheimer's disease (AD) is a progressive neurodegenerative disorder. It can cause a massive impact on a patient's memory and mobility. As this disease is irreversible, early diagnosis is crucial for delaying the symptoms and adjusting the patient's lifestyle. Many machine learning (ML) and deep learning (DL) based-approaches have been proposed to accurately predict AD before its symptoms onset. However, finding the most effective approach for AD early prediction is still challenging. This review explored 24 papers published from 2018 until 2021. These papers have proposed different approaches using state of the art machine learning and deep learning algorithms on different biomarkers to early detect AD. The review explored them from different perspectives to derive potential research gaps and draw conclusions and recommendations. It classified these recent approaches in terms of the learning technique used and AD biomarkers. It summarized and compared their findings, and defined their strengths and limitations. It also provided a summary of the common AD biomarkers. From this review, it was found that some approaches strove to increase the prediction accuracy regardless of their complexity such as using heterogeneous datasets, while others sought to find the most practical and affordable ways to predict the disease and yet achieve good accuracy such as using audio data. It was also noticed that DL based-approaches with image biomarkers remarkably surpassed ML based-approaches. However, they achieved poorly with genetic variants data. Despite the great importance of genetic variants biomarkers, their large variance and complexity could lead to a complex approach or poor accuracy. These data are crucial to discover the underlying structure of AD and detect it at early stages. However, an effective pre-processing approach is still needed to refine these data and employ them efficiently using the powerful DL algorithms.

ARTICLE HISTORY
Received: 13 Aug 2021
Revised: 04 Sept 2021
Accepted: 21 Oct 2021

KEYWORDS
Alzheimer's disease (AD)
Deep learning
Genetic variants
Machine learning
Mild cognitive impairment (MCI)
Normal control (NC)
Neuroimaging data
Single nucleotide polymorphism (SNPs)

INTRODUCTION

Alzheimer's disease (AD) is a fatal disease that slowly destroys brain's cells causing serious damages in the patient's body, mentally and physically. The symptoms start to appear gradually, starting from memory loss, confusion, and depression, and ending at losing the ability to eat and walk [1]. The prevalence of AD within the next 30 years is really shocking. A study conducted in 2013 to estimate AD prevalence in the United States from 2010 until 2050 revealed that the number of elderly people suffering from AD dementia will increase from 4.7 million to 13.8 million [2]. In fact, the continuous increase in the number of deaths due to AD dementia in the US has made it the fifth leading cause of death for people aged 65 and older [1]. Furthermore, the global impact of AD is more dreadful. According to the latest world health organization fact sheet for dementia statistics around the world, there are nearly 10 million new cases of dementia worldwide, and 60 to 70 % of them are caused by Alzheimer's disease. Extensive research has been conducted to discover the real reasons behind this mysterious disease and find the perfect cure that can impede this rapid increase in AD cases [3]. Unfortunately, an effective cure for AD has not been discovered yet [4]. However, as the cognitive impairment progressively increases, an early prediction of AD will greatly help reduce its impact through an early therapeutic intervention [5], and it will give the patient more time to adjust with its symptoms and improve their lifestyle [6]. Therefore, several machine learning and deep learning techniques have been proposed to detect AD at early stages. Nevertheless, proposing an optimal approach able to efficiently predict AD with high accuracy is still a big challenge.

Artificial intelligence (AI) is one of modern technologies that has been largely used in many applications to build intelligent systems that simulate human's way of thinking. Machine learning is a subset of Artificial intelligence. It was defined in 1959 by Arthur Samuel, a pioneer in AI and computer gaming, as "field of study that gives computers the ability to learn without being explicitly programmed". Machine learning algorithms could overcome the static program instructions and create computational models able to automatically learn from data and derive different decisions and predictions [7]. There is a wide variety of ML algorithms that were successfully used in many fields such as healthcare, marketing and education [8] . However, with the advent of big data, deep learning, which is a subset of machine learning, has remarkably surpassed traditional methods [9]. DL algorithms have achieved high levels of accuracy in many areas such as voice and face recognition [10].

*CORRESPONDING AUTHOR | Ghada M. Fadhl Alqubati | ✉ ghada.alqubati101@gmail.com

© The Author(s) 2021. Published by Universiti Malaysia Pahang Al-Sultan Abdullah Publishing. This is an open access article under the CC BY-NC 4.0 International license

Machine learning algorithms have been increasingly utilized to analyse medical data and extract several features that can be used to understand many aspects related to the disease such as the disease pathology and human brain malfunctions [11]. And with the current progress in machine learning technology, new techniques have been developed to predict AD and model its progression [12], among which supervised machine learning algorithms have proven their efficiency to learn from a massive amount of data within a very short time, and demonstrated their capability of helping doctors to accurately predict diseases at early stages [13]. However, with the extensive breakthrough in neuroimaging technologies resulting in high complex and large-scale data, deep learning technology has intriguingly exhibited its preference over traditional ML methods at interpreting neuroimaging data with high dimensionality and precisely detect AD [14]. In fact, it was recently proven that DL technology has become the foundation for the prediction of AD [15].

Although ML and DL have achieved high precision at detecting AD with neuroimaging data, most of these approaches could lack the ability to discover the susceptibility of the disease early enough [16]. Therefore, some researchers have supported their analysis by including genetic data to other neuroimaging modalities [17]–[19] . This is because AD is considered as a complex disease with a genetic basis [20]. The complex nature is derived from the many factors that can contribute to the disease such as environmental factors and genetic or inheritance factor [21], in which the latter plays a fundamental role in the disease pathogenesis as it may contribute to 70% of risk factors [22].

According to age pattern, AD onset was divided into two subsets: early onset AD (EOAD) and late onset AD (LOAD). The first set affects people aged less than 65 and has 5% of total AD cases, whereas the second one has 90% to 95% of total cases and affects elderly people aged more than 65 [23]. EOAD is less complicated and more understandable than LOAD, in which many genes can be associated with it [21], [23]. Hence, many technologies have emerged to understand and decode the human genome and turn it into a readable format so researchers can study it closely and extract vital genomic biomarkers [24], [25]. These markers can enrich the knowledge about LOAD characteristics and its etiology, and lead to an early diagnosis and therapy development [26].

Machine learning algorithms with their powerful abilities at manipulating multi-dimensionality data have proved their excellence at increasing prediction precision of complex diseases using genetic markers (usually known as SNPs) [27]. They have been involved in many approaches and helped discover many disease genes associated with AD dementia.

The wide adaptability of machine learning technology into the health sector has resulted in a broad range of available medical datasets for researchers [28]. In fact, a great deal of open access data repositories and a wide range of medical datasets can be easily accessed such as massive electronic health records, neuroimaging datasets, and genomics biomarkers.

In this survey, we explored 24 papers from 2018 till 2021 based on machine learning and deep learning techniques to early predict AD. We gathered the latest AD prediction approaches and divided them into ML based and DL based approaches. We further classified them based on the type of medical data, and discussed their workflow and results to identify their advantages and disadvantages. The review summarized an overall knowledge about the recent ML and DL technologies and their findings in the context of AD early prediction. It provided researchers valuable insights into research gaps and future research.

The rest of the paper is organized as follows: section 2 is for the review methodology; section 3 states some types of AD biomarkers; section 4 explores a number of recent machine learning and deep learning based approaches for detecting AD. This section is divided into two subdivisions: Machine learning based approaches, and deep learning based approaches. In each division, the approaches are separated into a number of categories based on the data type used in the method. Section 5 is for findings discussion and results comparison. Finally, a conclusion and future work are presented in section 6.

MATERIALS AND METHODS

The continuous advances in machine learning and deep learning technologies, and the large diversity of biological and medical data have opened the way for a large field of various research studies for AD classification and prediction. In this review, we focused on 24 papers published from 2018 and 2021. These papers were selected to explore the recent findings in AD prediction using machine learning and deep learning algorithms on various biomarkers. We firstly outlined some types of AD biomarkers, demonstrated in Fig 1. Then, we objectively summarized the selected papers by dividing them into two main categories based on the type of AI learning technique used: approaches using ML algorithms and approaches using DL algorithms, demonstrated in Fig 2. Each category was split based on the type of data exploited in the approach. In the first category, the approaches were split into six sub categories: images data, large scale health data, gene expression data, genetic variants data, mobility and cognitive data, audio data. In the second category, the approaches were split into four sub categories: images data, large scale health data, genetic variants data, and heterogeneous data. We explained each approach in terms of its workflow, algorithms, data type, and performance results. After that, we discussed all approaches from different perspectives, outlined their pros and cons, and briefly compared their findings using area under the curve (AUC) and accuracy (ACC) as the evaluation metrics since these two metrics were the common metrics used in all papers. Lastly, we summarized all of them in two tables based on the dataset name, data type, algorithm/s, evaluation technique, and testing results.

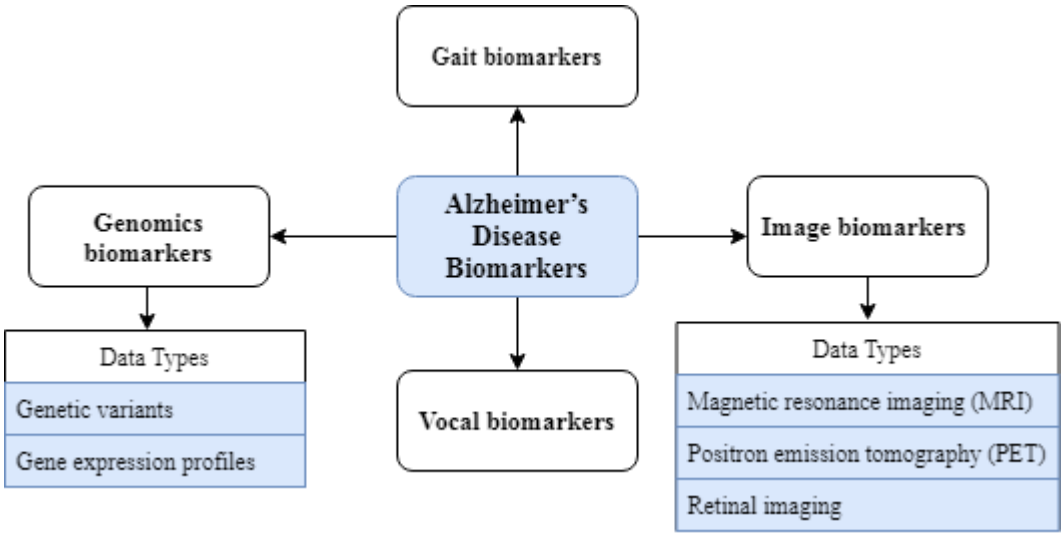

Figure 1. AD biomarkers categories.

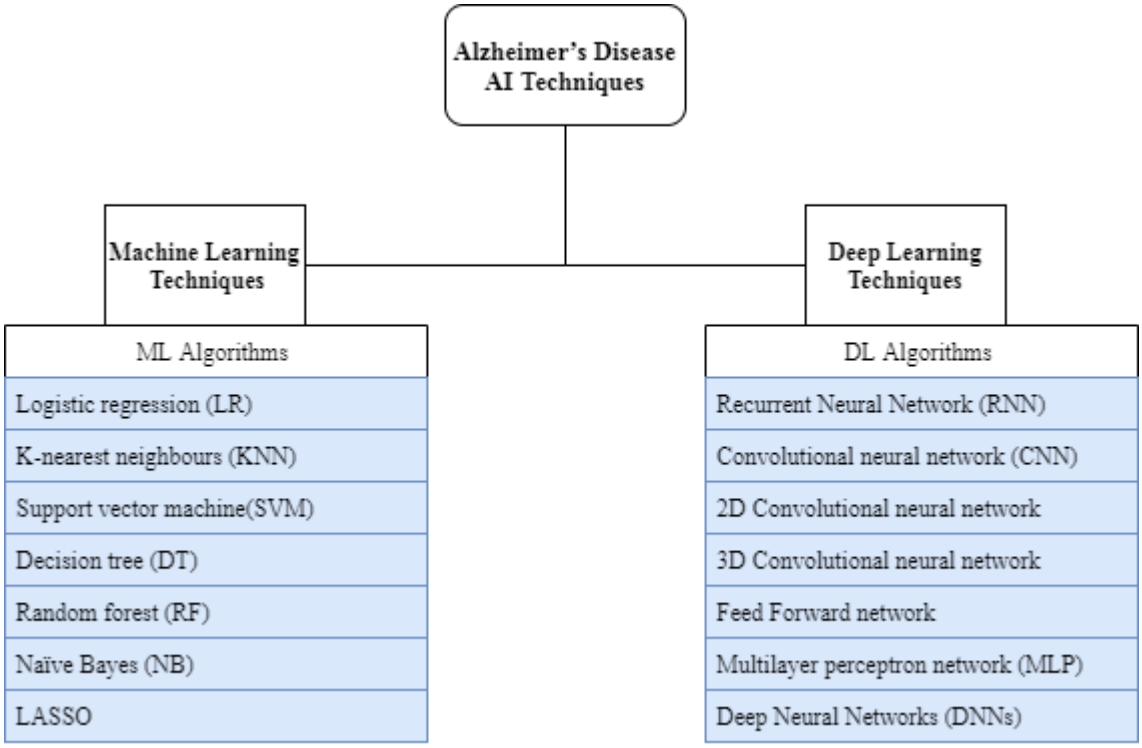

Figure 2. AD ML and DL categories.

BIOMARKERS OF ALZHEIMER'S DISEASE DETECTION

AD biomarkers have played an integral role in understanding its structure and monitoring its progression to help its early detection and treatment [29], [30]. This section summarizes some types of these biomarkers.

Image biomarkers

Many imaging techniques have emerged for AD diagnosis such as structural magnetic resonance imaging (sMRI) scans, shown in Figure, which is one of neuroimaging techniques used to support doctors' diagnosis of AD and help measure the size of degeneration of some brain areas that can lead to early detection [31]. sMRI scans , shown in Fig 3, produce high contrast images used to measure the volume of the grey and white matter of the patient's brain [32]. They can detect the atrophic alterations in the human brain that could lead to severe damages causing AD [33]. Another neuroimaging technique used for detecting AD is amyloid and tau positron emission tomography (PET). It was proven that the accumulation of amyloid and tau, which are some types of proteins, in the brain can severely damage its cells and

lead to AD [34]. Amyloid and tau PET biomarkers have become increasingly important for studying the abnormal accumulation of these proteins and understanding disease pathology [35]. There are many open access databases for sMRI and PET such as Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets [36] and Open Access Series of Imaging Studies (OASIS) dataset [37]. Furthermore, as MRI and PET could be limited in accessibility and expensive [38], [39], a new non-invasive, inexpensive technique known as retinal imaging, shown in Fig 4, has recently been used as AD biomarkers. These biomarkers show the abnormal changes of retinal vascular that could be associated with AD [38]. An example of retinal images databases is UK biobank [40].

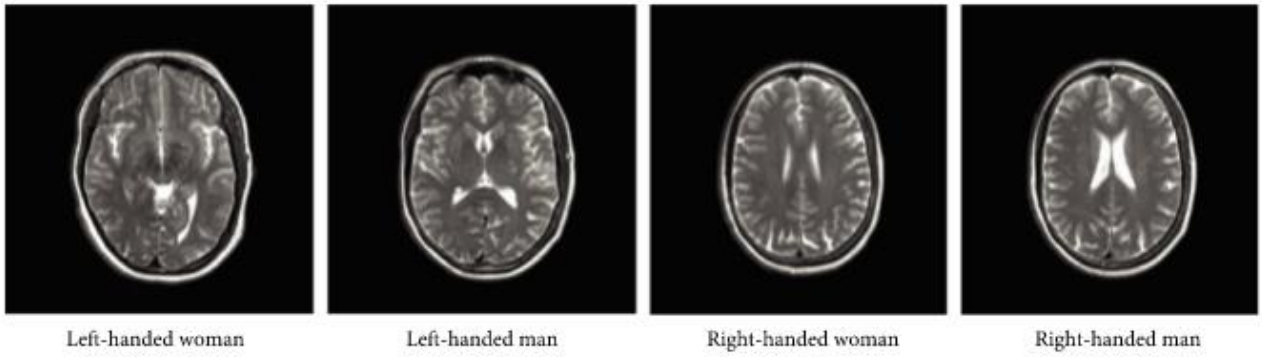

Figure 3. 2D Structural MRI examples of AD patients [41].

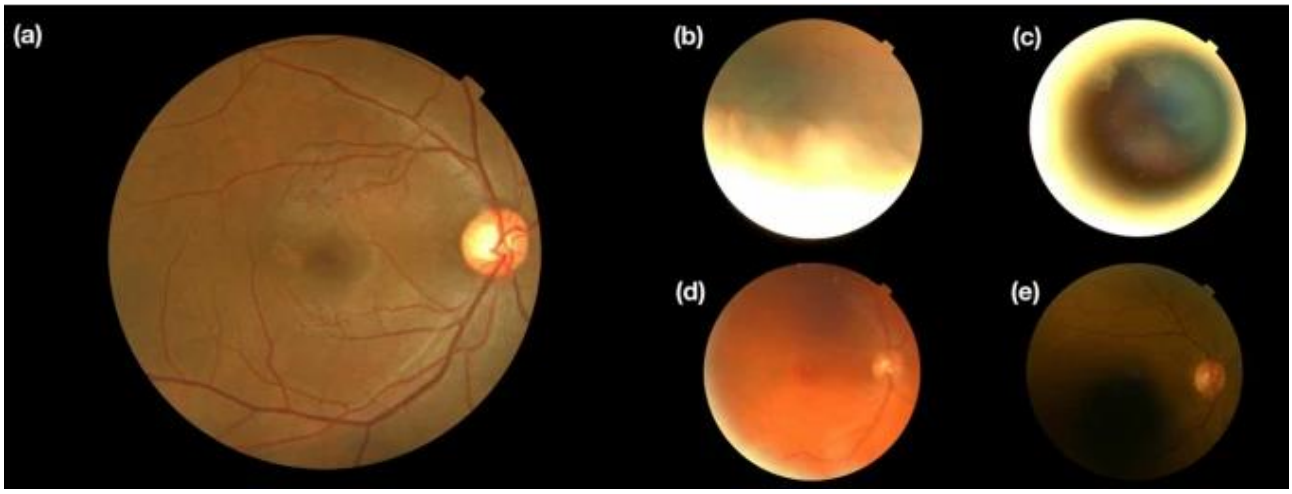

Figure 4. Retinal image with different degree of quality [38].

Genomics biomarkers

With the advance of DNA sequencing technologies and with the help of Genome wide association studies (GWAS), many disease-associated genes to AD have been discovered [21]. Next generation sequencing (NGS) is a DNA sequencing technology. It is a cost-effective, highly accurate deep sequencing technique that can sequence the whole human genome into millions of four-letters sequences within only one day [24]. Various platforms using NGS technology have paved the way for a wide range of studies to explore different regions of the human genome and discover many genetic variants contributing to complex diseases [42]. One of those studies is GWAS that concerns analysing human genetic variations to define the genetic risk factors of a complex disease [43]. Genetic variants can be a single alteration in DNA sequence known as single nucleotide polymorphism (SNPs) or longer alteration such as insertion and deletion variations (indels) and copy number variations (CNVs) [44] . Moreover, gene expression is another type of genomics biomarkers. Gene expression is the set of instructions encoded in DNA and used to build protein molecules (gene products) [45]. DNA microarray is one of many technologies used for gene expression profiling [25]. It is a powerful tool that can monitor the expression of thousands of genes at the same time and profile valuable information about the gene expression process. Gene expression profiles can help understand the basic genetic structure of a disease through discovering genes involved in its formation [46]. They have the ability to visualize the physiological changes of an AD patient and guide many researchers to understand the biological aspects of the disease pathology [47]. There are a wide range of genome datasets such as ADNI [36], and Dementia and Traumatic Brain Injury (TBI) Study. ADNI provides two resources of genetic variants: GWAS genetic variants and Whole genome sequencing (WGS) dataset. In the WGS

dataset, genetic variants are stored in a variant call format (VCF), as shown in Fig 5. VCF is a standard representation of genetic variants including SNPs, indels and other structural variants [48].

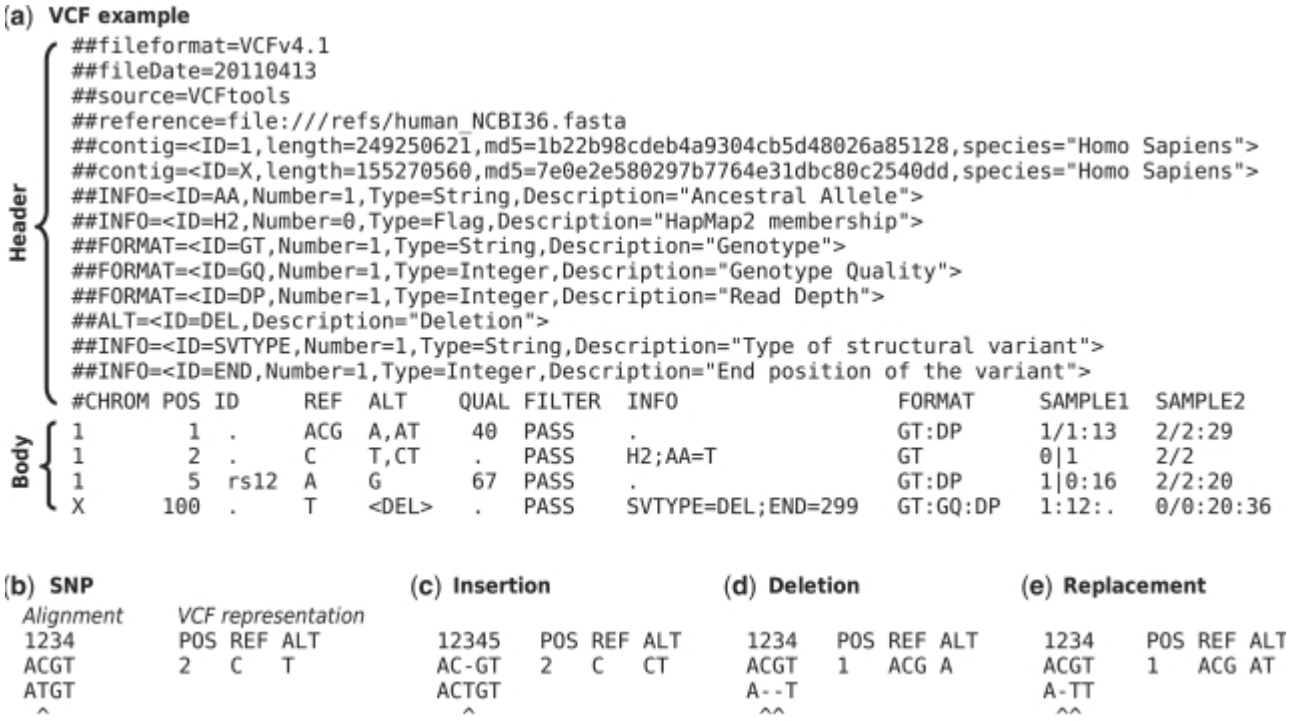

Figure 5. An example of VCF file [48].

Vocal and Gait biomarkers

Vocal biomarkers can be collected in non-invasive and inexpensive manner, and they can be used to analyse audio segments of subject's speech and extract risk features associated to AD [49], whereas, gait or walking biomarkers can be used to monitor subject's movements and reactions to extract risk features associated to AD [50], as shown in Fig 6.

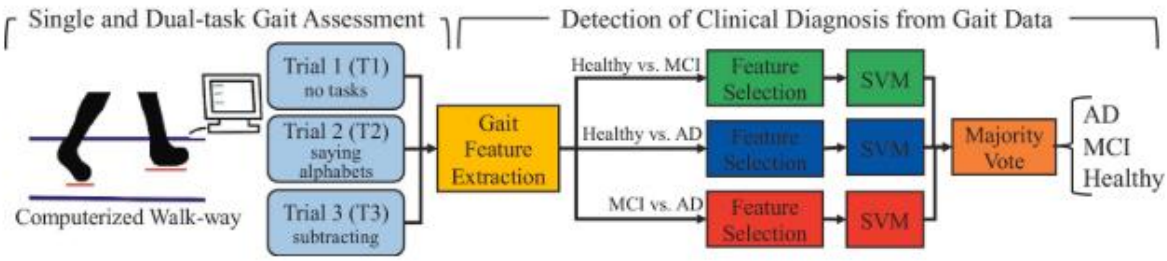

Figure 6. A proposed approach using gait data to diagnose AD [50].

ML AND DL-BASED APPROACHES FOR ALZHEIMER'S DISEASE DETECTION

In this section, the latest approaches proposed to predict AD dementia at early stages are illustrated. They are divided into two main parts: approaches with machine learning techniques and approaches with deep learning techniques. For each part, the approaches are split based on the type of biomarkers used.

Machine learning-based approaches

This section delineates recent machine learning-based approaches based on the type of data used to predict AD.

Medical imaging data

There is a variety of imaging data used by machine learning algorithms such MRI scans data, PET scans data, and retinal images data. In the following, ML approaches using two types of imaging data: MRI and retinal images data were explored:

Magnetic resonance imaging (MRI) data

In [51] longitudinal structural MRI (sMRI) data of 150 participants were used for AD classification through a 4-stage automated pipeline. The first stage was for data pre-processing. The second stage was for dividing data into two main sets: training set and test set. The main training set was also divided into three subsets: train, validate, and test sets. In the third stage, the three subsets were used by 17 supervised machine learning algorithms to build predictive models, and the best model resulting from this stage was tested in the final stage using the main test set. The best algorithm was Random forest (RF) with AUC of 0.8722. Another approach in [52] employed structural MRI for differentiating people with AD dementia from people with vascular dementia. Researchers used a collection of MRI scans for 58 subjects with AD and 35 subjects with VD. The approach went through multiple steps of image pre-processing such as skull-stripping and alignment in order to select the appropriate features for the training stage. Four machine learning algorithms were applied: Support vector machine(SVM), K-nearest neighbours(KNN), RF, and Logistic regression (LR). The SVM with radial basis function (RBF) kernel attained the best outcomes with AUC of 0.861. Moreover, Researchers in [53] utilized 1,167 sMRI scans to classify normal cognitive (NC) state and three different states of dementia: early MCI, late MCI, and probable AD. The approach trained six ML classifiers: KNN, Decision tree, RF, Naïve Bayes (NB), linear SVM and non-linear SVM with RBF kernel. The testing results showed that non-linear SVM with RBF kernel accomplished the best classification performance for all stages with AUC of 0.76.

Retinal vasculature imaging data

A recent study in [38] exploited retinal biomarkers to predict AD through a machine learning pipeline. The pipeline consisted of three stages. The first stage was for selecting images with sufficient quality, the second stage was for generating vessel maps and using T-test for feature selection, and the final stage was for model building using SVM classifier. The classifier demonstrated an overall accuracy of 0.824.

Large scale health data

As machine learning algorithms have exhibited their superiority in big data, many approaches have used them to analyse a massive amount of health data and extract important features for predicting AD. The researchers in [54] developed a number of predictive models for predicting definite AD and probable AD within 4 years. They applied three machine learning algorithms, RF, SVM and LR on large-scale data including clinical tests, participants and family information, and prescribed medications. RF classifiers surpassed other classifiers in 1-year to 4-year prediction of definite AD in which results ranged from AUC of 0.775 to AUC of 0.677. In addition, in [55] researchers employed an extensive data collection of clinical tests, neuropsychological tests, social and demographic information to predict the conversion of a patient from mild cognitive impairment (MCI) to AD dementia within three years. By using weighted rank average ensemble technique, they built an ensemble ML model consisting of 13 supervised machine learning algorithms such as KNN, LR, RF and NB, and achieved a performance of AUROC 0.88. Another approach in [4] used a large scale of health data for early prediction of AD. They collected multiple attributes from different tests such as Mini-Mental state examination, clinical dementia rating, estimated total intracranial volume, and other information of participant's socioeconomic status and education background. They utilized a number of machine learning classifiers to train and validate their models such as RF, LR, NB, SVM with linear kernels, in which the latter demonstrated best accuracy of 0.95.

Gene expression profiles data

In [46] gene expression data were exploited to classify AD and discover new genomics biomarkers associated with AD. The researchers at first ranked expressed genes with P-value by using T-test in order to remove genes with P-value less than 0.5 as they have significantly different expressed values than the two sample classes, AD, and NC. After removing differentially expressed genes, 2000 genes were selected for training and testing, and five machine learning classifiers were employed. The best classifier was SVM with a linear kernel. On the other hand, three techniques were utilized for feature selection: Principal component analysis (PCA), RF, and Extra tree classifier. After analysing the extracted features or genes, the 9 genes selected by PCA were chosen and joined with the overlap set of genes selected by the three methods. The new set of 14 genes were tested by the SVM classifier, since it got the best classification results, and the results were better. The new set of genomics biomarkers was considered as an influential set associated with AD. In contrast, the approach in [56] used differentially expressed genes (DEGs) extracted from four regions of the human brain to study their connection with the disease as researchers believe that these kinds of genes coming from different regions are correlated to AD. They started by removing redundant data from each sample since gene expression data were taken for four regions of the same person. Then, the expressed genes were ranked with P-value by using linear mixed effect model (LMM) technique. After that, genes with minimum P-value that were differentially expressed were enriched by gene ontology to explain their biological implications. Lastly, all genes from both classes, AD and non-AD, were enriched by using a gene ontology database in order to find the functional connection or pathways between them and DEGs. Top ten and top six of DEGs were chosen and tested by four ML algorithms, in which RF algorithm achieved the best accuracy of 0.73 and 0.83 respectively.

Genetic variations

In [57] researchers used SNPs data for classifying AD and extracting the genetic variants associated with AD. They suggested a new approach to improve the classification accuracy by using the misclassified samples. At the beginning, they trained three ML classifiers: BSWiMS, GALGO, LASSO, and from the best classifier, LASSO, they selected the misclassified testing samples. Then, they extracted the related SNPs of these samples and retrained the model with the LASSO classifier. After that, they merged the features extracted from all samples and the features extracted from the misclassified samples, and used them to train the model. The results achieved by using the last set of features demonstrated the best testing performance with AUC of 0.842. One more approach utilized SNPs data in [16]. The researchers designed an ensemble model to predict AD. They first pre-processed genetic variants by applying quality control procedures. Then, they picked the top 2,500 SNPs to build the ensemble model consisting of five ML classifiers by using a benchmarking tool called feature selection algorithm for computer aided diagnosis (FRESA.CAD). After validating and testing the models, the classifiers performance ranged from AUC of 0.6 to AUC of 0.7, whereas the ensemble model achieved a better output with AUC of 0.719. However, when the ensemble model was trained with the top 1000 SNPs, it attained a result with AUC of 0.554. Moreover, the ensemble model resulted in eight genes that were the most selected genes among all classifications, and these genes were known for their strong association to AD. Another approach suggested in [58] is to study and discover the effect of genetic mutations related to AD through extracting the most influential features and using them to segregate the harmful SNPS from harmless SNPs. In the suggested approach, a two-stage feature selection was applied to select the most important features. In the first stage, recursive feature elimination cross validation (RFECV) was used to select 39 features. In the second stage, forward feature selection was used to select the best feature combination. After selecting the best combination of 11 features, a model was trained on these properties using a random forest algorithm, and the achieved result was AUROC of 0.8949.

Mobility and cognitive data

Researchers in [50] used dual-task gait assessments data for classification AD, MCI, and NC. The gait features were extracted from a gait analysis software with a pressure sensitive carpet. The subjects underwent dual-task valuations through walking and testing their cognitive ability such as memory, language and attention at the same time. SVM classifier was trained on the data, and it achieved an average accuracy of 0.78

Audio data

A new approach was suggested in [49] to employ speech data for predicting AD at early stages. In the approach, the audio data were gathered and divided into 1-second segments. After that spectrogram features were extracted and used to train ML models using five ML algorithms. The classifier's performance was tested on two data sets, in which logistic regression CV classifier achieved best results with accuracy of 0.833 and 0.844 in the two datasets.

Deep learning-based approaches

This section outlines recent deep learning-based approaches based on the type of data used to predict AD.

Medical imaging data

Magnetic resonance imaging (MRI) data

Researchers in [41] employed a two dimensional convolutional neural network (2D CNN) to predict AD. They used MRI data to train their model in which they tried a number of inputs for the last hidden layer ranging from 120 inputs to 130 inputs with a dropout rate ranging from 0.1 and 0.5 in order to get the best performance, which was found at 121 units with a drop rate of 0.2. The model attained a testing accuracy of 99.30. Another approach exploited MRI data in [59] to predict AD. Researchers used structural MRI features extracted from the hippocampus area of 933 subjects. They designed a lightweight three dimensional CNN by using the deep visual attributes extracted from another model called 3D Dense CNN and the global shape attributes extracted from hippocampus segmentations. The features then were combined in a fully connected layer followed by a softmax layer. The model accomplished an accuracy of 92.52.

Positron emission tomography (PET) data

Raeserchers in [60] used amyloid or tau PET features for AD classification. At first, they trained 3D CNN for classifying AD and NC. Then, they used the trained model to predict the conversion of MCI state to AD state, in which a subject with a probability close to 1 was classified as an AD conversion, whereas a subject with a probability close to 0 was classified as nonAD conversion. After that, a layer wise relevance propagation (LRP) algorithm was used to extract features resulting from the model to be visualized in a heat map to show brain areas closely related to AD. The average accuracy of classifying AD and NC was 90.8.

MRI data + PET data

The approach in [61] employed two image modalities: MRI scans and amyloid PET scans to predict AD. After pre-processing both modalities, two identical CNNs of the two modalities trained on the same time. The weights of both

networks were merged at the last hidden layer consisting 128 inputs to form a fused network with one output layer. The testing results of this network was with accuracy of 92.34.

Large scale health data

In [62] researchers used longitudinal electronic health records from 2007 to 2017 including many features such as subject's age, background and clinical test results. Three models were trained on these data to predict MCI and AD within three to eight years using recurrent neural network (RNN), RNN with trained weights of another model, and a feed forward network. In the latter, researchers inserted three features, sex, age, days of collecting data, directly to the last hidden layer to ensure that all of their weights are included. The best results ranged from 0.81 to 0.84.

Genetic variations data

In [63] researchers exploited SNPs data only to predict AD. They used whole genome sequencing data of 42,908,833 SNPs. After applying a quality control pipeline to remove bad SNPs, they used 1,884 SNPs for building their predictive models. They suggested two neural network architectures, DNN and 1D CNN. For evaluating the performance, they divided the SNPs into a number of subsets based on their p-value that was copied from the international genomic of Alzheimer's project (IGAP) report. The best performance was for DNN on a subset of 200 SNPs with AUC of ≈ 0.62.

Heterogeneous data

Some approaches have used different types of biomarkers in order to improve the prediction accuracy either by merging them into one unified form, or by using them separately and merging final results.

Neuroimaging + Genetic variants data

Researchers in [17] proposed to merge SNPs data with brain region of interests (ROIs) data. This is because they believe that this kind of data can directly describe the disease, whereas genetic biomarkers can describe its etiology, and as a result, the neural network could fail when dealing with these biomarkers only. Thus, they assumed that the structural information of brain regions can help the network understand genome data and improve its accuracy. At first, SNPs and ROIs information were normalized and ranked based on their degree of importance by using random forest algorithms. After merging them, the total number of features was 542 features. After that, a deep learning model was trained and tested on these features, and the results showed an improvement in the network performance. The best result was for the top 10 SNPs and ROIs with AUC of 0.80. Another approach employing images data and SNPs data was suggested in [18]. The researchers suggested a method to improve the accuracy of a conventional neural network (CNN) used to predict AD by merging its predictions with another network's predictions. In the approach, the two networks, CNN with MRI data and multilayer perceptron network (MLP) with SNPs data, were trained separately. After getting the output of both networks, an ensemble gate merged them to form the final prediction result if the prediction accuracy of CNN was low. Otherwise, the final result would be for CNN prediction only. After the approach evaluation, the prediction accuracy for 75 subjects improved from AUC of 0.9232 when using MRI scans only to AUC of 0.936 when using both MRI and SNPs data. Furthermore, in [19] researchers used MRI data and SNPs data of APOE ε4 allele and 19 SNPs known for their strong contribution to AD. They suggested merging MRI features and SNPs features and building predictive models to predict the conversion from MCI state to AD state. They trained 100 models using DNN and 100 models using logistic regression (LR). The models were trained to classify AD and NC states. Then, they were tested to predict MCI conversion. The DNN showed better performance than LR with AUC of 0.835.

DNA methylation profiles + gene expression profile

DNA methylation is a process involved in gene expression regulation [64]. Its potential effect is usually at DNA regions known as CpG islands. Some studies believe that there is a correlation between gene expression and DNA methylation. Hence, we found that researchers in [65] used both of them to predict AD. They used gene expression and DNA methylation profiles extracted from the prefrontal cortex. As both profiles cannot be merged directly because of their different behaviour and characteristics, the researchers proposed a feature selection method to extract features from both profiles into two features, one for genes and the other for CpG probes. The method had two steps. The first step was for filtering differentially expressed genes (DEGs) and differentially methylated positions (DMPs). As every DMP has its related genes, researchers in the second step merged both features by intersecting genes that were differentially expressed and differently methylated as they believe that these genes have a strong connection to the disease. After that, DNN was built and optimized with Bayesian hyper-parameter optimization, and the model achieved an accuracy of 82.3% and AUC of 0.797.

DISCUSSION

This section discusses all approaches from different perspectives to draw implications about their results, strengths and limitations, and make recommendations for future work.

Few machine learning research studies have tried to use new types of data other than Neuroimaging and genetic variants, which are the most modalities used by ML methods, to predict AD. One recent study [50] that analysed gait movement and patient's cognitive responses to extract features capable of classifying AD patients from a cognitively normal person. Analysing such data with machine learning technology could greatly contribute to discovering the subtle

cognitive or physical changes that a patient may exhibit long time before AD onset, and this will help doctors to discover the disease early enough. However, as the utility of these biomarkers is still limited by few research studies, further research might be needed to assure their significance. Another study used Audio data to predict AD [49]. These data are inexpensive and more accessible compared to other modalities such as Neuroimaging data. And same as physical and cognitive features, some Audio features could be an important indication of AD development in future, and they could help doctors and patient's families to predict AD susceptibility from the way this patient talks or sounds. Nevertheless, further research might be also required to explore their relevance to AD early prediction.

Moreover, although DL technologies have shown a higher precision performance than ML technologies [66], it was found that the number of ML based approaches employing genetic variants only to predict AD were more than DL based approaches. In fact, only one research was found using SNPs data as the only modality to classify the disease. This might be because of the complicated nature of these kind of features by which a neural network usually achieves poor classifying accuracy. This could also be due to the limited number of samples compared to the enormous number of features in most SNPs datasets that might affect the network performance because DL algorithms require huge amount of instances [33]. Therefore, most researchers have tended to use ML algorithms for feature selection and classification. Another reason could be due to the limited number of samples compared to the enormous number of features in most SNPs datasets. This may also affect the network performance because DL algorithms require a huge amount of instances. On the other hand, many DL based approaches have involved genetic variants with other modalities [17]–[19], mostly neuroimaging modalities, as a way to improve network performance. And in spite of adding more complexity to their approaches, most of them got only an accuracy improvement of 2% to 3%. As genome biomarkers such as genetic variants play an inarguable role at understanding the disease's underlying structure [26], and because of the promising capability of DL technology with genetic data [67], further research on employing this technology with genetic variants could help explore them more deeply and define the vital regions in human DNA that are strongly related to AD development. However, an effective pre-processing and quality control pipeline could be the decisive step for reducing the complexity and variety of SNPs data, and leading to a noticeable improvement in network performance.

In addition, when the results of ML based approaches were compared in terms of genetic variants data and neuroimaging data, shown in Table 1 and Fig 7, it was noticed that they were relatively close with an average AUC of 0.82. Nevertheless, the results were largely different in DL based approaches, shown in Table 2 and Fig 8, in which methods using neuroimaging data achieved an average ACC of 93.74, while the others using SNPs data achieved an average AUC of 0.67.

Table 1. ML approaches results with MRI and SNPs data.

Ref No.	Data type	AUC
[51]	sMRI	0.8722
[52]	sMRI	0.861
[53]	sMRI	0.76
[57]	SNPs (482)	0.842
[16]	SNPs (2500)	0.719
[58]	SNPs (11)	0.8949

Table 2. DL approaches results with MRI, PET and SNPs data.

Ref No.	Data type	*ACC/AUC
[41]	MRI	0.993
[59]	sMRI	0.9252
[60]	amyloid PET	0.908
[61]	amyloid PET + MRI	0.9234
[63]	SNPs (200)	0.62
[17]	SNPs (20 & 50)	0.68
[18]	SNPs (41)	0.6807
[19]	SNPs (20)	0.689

*Note: images data results were measured by ACC

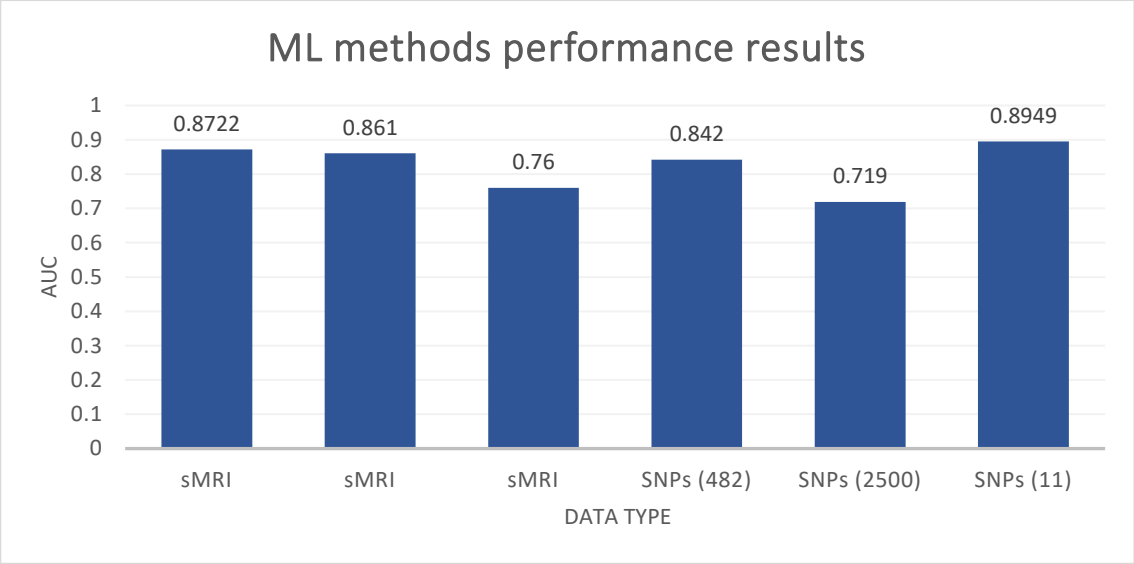

Figure 7. ML approaches results comparison in terms of MRI and SNPs data.

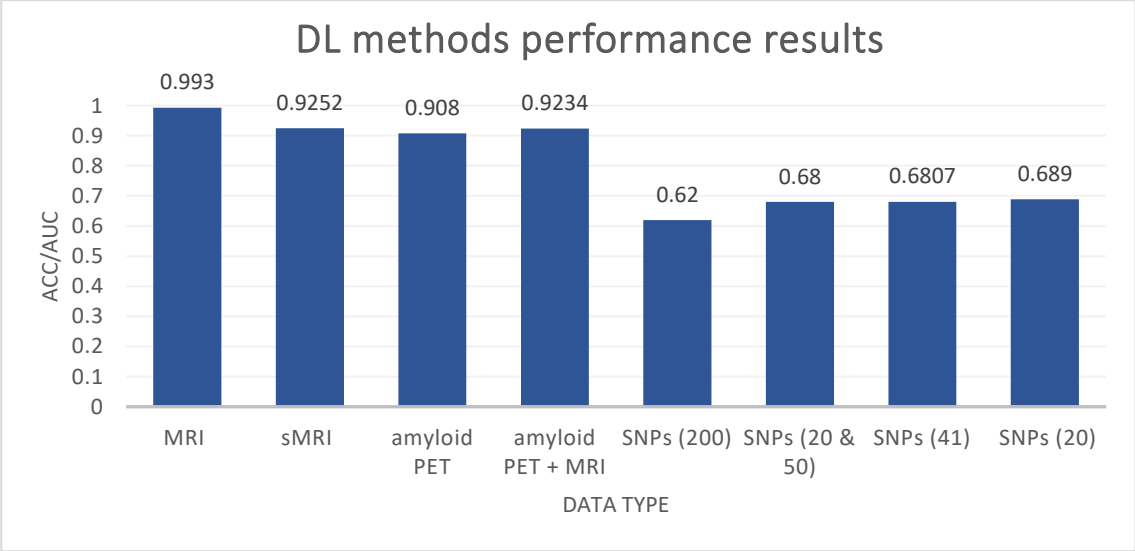

Figure 8. DL approaches results comparison in terms of MRI and SNPs data.

Furthermore, when the results of ML and DL based approaches were compared in terms of genetic variants data and neuroimaging data, shown Table 3, it was found that DL based approaches had achieved better performance in Neuroimaging data compared to ML based approaches, while they were relatively poor with SNPs data.

Table 3. ML & DL approaches average results.

ML/DL	Data type	AUC/ACC
ML	Neuroimaging data	AUC= 0.825
DL	Neuroimaging data	ACC= 0.937
ML	Genetic variants data	AUC= 0.818
DL	Genetic variants data	AUC= 0.667

Additionally, Table 4 demonstrates strengths and limitations of ML and DL based approaches in terms of data type, and Table 5 and 6 demonstrate a summary of all approaches mentioned in the survey in terms of the algorithms used, dataset name, modality type, evaluation technique, and results. They provide a comparative analysis that helps formulate a knowledge about the latest accuracy level of AD prediction, and the type of modalities and algorithms related to that accuracy.

Table 4. ML & DL based approaches strengths and limitations.

Technique	Data type	Strengths	Limitations
Machine learning	Images data	Some of open access datasets are freely available, and the technique achieves very good classification accuracy	Data acquisition is difficult and expensive, and AD susceptibility prediction at early stages could be poor as genetic factors, which form 70% of the risk factors, are not considered.
	Large scale health data	Data acquisition is cost effective, and a wide range of open access datasets are available. And the technique achieves an excellent accuracy	AD susceptibility prediction at early stages could be poor as genetic factors are not considered
	Gene expression data	Data acquisition is cost effective, and a wide range of open access datasets are available. Besides, 70% of risk factors are considered, and the technique achieves very good classification accuracy	The data requires a long pre-processing and preparation process. And the technique applies feature selection for only small subsets of genes to achieve higher accuracy
	Genetic variants (SNPs) data	Data acquisition is cost effective, and a wide range of open access datasets are available. Besides, 70% of risk factors are considered, and the technique achieves very good classification accuracy	The data need an efficient quality control (QC) pipeline to extract good SNPs. And only small subsets of genetic variations are considered to achieve higher accuracy
	Mobility and cognitive data	The technique achieves good accuracy, and its extracted risk factors could help doctors and patient's family to detect the diseases at early stages	Although these data can be easily adapted in clinics, they are still limited and few research studies have explored them to predict AD. Therefore, the extracted risk features might be still indecisive for AD early detection especially that the brain and genetic variations are not considered
	Audio data	Data acquisition is inexpensive, and non-invasive. The technique achieves very good accuracy, and its extracted risk factors could help doctors and patient's family to detect the diseases at early stages	Although these data are non-invasive and easily collected, they are still limited and few research studies have explored them to predict AD. Therefore, the extracted risk features might be still indecisive for AD early detection especially that the brain and genetic variations are not considered
Deep learning	Images data	Some of open access datasets are freely available, and the technique has great ability to analyse raw data without prior feature selection and achieves excellent classification accuracy	Data acquisition is difficult and expensive, and the technique is computationally expensive as it requires hardware with high computational capabilities to minimize training time. Besides, AD susceptibility prediction at early stages could be poor as genetic factors are not considered
	Large scale health data	Data acquisition is cost effective, and a wide range of open access datasets are available. The technique can rapidly analyse huge amount of raw data and achieves very good accuracy	AD susceptibility prediction at early stages could be poor as genetic factors are not considered.
	Gene expression & DNA methylation data	Data acquisition is cost effective. Besides, 70% of risk factors are considered and the technique achieves very good classification accuracy	These data are heterogeneous and require multiple steps to pre-process them and merge them. Very few studies have considered this combination. Therefore, its extracted risk features might be still uncertain, and further research is needed.
	Genetic variants data	70% of risk factors are considered, and the technique	It requires a huge amount of data and an efficient QC pipeline for extracting good

		can handle larger sets of genetic variants and automatically recognize related patterns without prior feature extraction	SNPs to avoid overfitting and achieve high accuracy. However, most of the available datasets are highly complex and variance with a huge number of features and limited number of samples. Besides, very few studies have used this modality without merging it with other modalities and achieved a relatively poor accuracy. Therefore, further research is needed.
	Images & genetic variants data	A wide range of research studies have used this combination and achieved an excellent accuracy	The approach is highly complex and computationally expensive, and data acquisition is difficult and expensive

Lastly, as genomics biomarkers form 70% of risk features, using them for predicting AD would be essential. So far, there have been many approaches using ML algorithms with genetic data, and most of them achieved good classification accuracy. Nevertheless, few approaches have used genetic data such as SNPs data with DL algorithms, and have poor accuracy. In fact, it was found that neuroimaging data and genetic variants were the most utilized modalities by DL technology. However, using genetic variants only to predict AD is still limited, and further research is needed.

Table 5. A summary of Machine Learning approaches.

Ref No	Dataset/s	Data type	Algorithms	Evaluation technique	Best performance metrics	
					Area Under the Curve (AUC)	Accuracy (ACC)
[51]	Open Access Series of Imaging studies (OASIS)	longitudinal sMRI scans	17 SML such as SVM, RF, Decision tree (DT), Stochastic gradient Descent (SGD)	10-fold CV with 10 iterations	0.8722	86.84
[52]		Structural magnetic resonance imaging (sMRI)	K-nearest neighbours (KNN), LR, RF, SVM	Dataset:70 % Training, 30% testing/ 10-fold CV	0.861	84.35
[53]	ADNI	sMRI	DT, linear SVM, nonlinear SVM with RBF kernel KNN, NB, RF	10-fold CV	0.76	75
[38]	open Access UK Biobank	retinal vasculature imaging data	SVM	5-fold CV		82.4
[54]	Korean National Health insurance service dataset	large amount of administrative health information	Logistic regression (LR), Random Forest (RF), Support vector machine (SVM)	Nested 5-fold stratified cross validation with five cycles	1-year: Definite AD: 0.78, Probable AD: 0.76 2-year: Definite AD: 0.73, Probable AD: 0.69 3-year: Definite AD: 0.68,	

					Probable AD: 0.64	
					4-year: Definite AD: 0.73, Probable AD: 0.68	
[55]	Alzheimer's Disease Neuroimaging initiative (ADNI)	a collection of information of clinical and cognitive tests and subject's background	13 Supervised Machine learning (SML) such as LR, RF, SVM, Naïve Bayes (NB), Gradient tree boosting	Stratified 10-fold CV	0.88	
[4]		large-scale of health records including clinical, personal, and cognitive information	DT, SVM, LR, NB, RF	5-fold CV		95
[46]	GSE5281 brain DB	Gene expression data of 24,438 genes, 87 cases (AD) & 74 controls (NC)	SVM with linear kernel, SVM with Gaussian kernel, NB, DT, RF	2/3 of dataset for training & 1/3 for testing		2000 genes: 89.80
						14 genes: 93.9
						9 genes: 93.9
[56]	RNA sequencing data downloaded from TBI study	Gene expression data of 50,281 genes, 15 cases (AD) & 30 controls (NC)	SVM with linear kernel, SVM with radial basis function (RBF) kernel, RF, Qbayes	Dataset:70 % Training, 30%		top 10 genes: 73
						top 6 genes: 83
[57]	GWAS	620,901 SNPs & 5,220 subjects	BSWiMS, GALGO, LASSO	Dataset:80 % Training, 20% testing/ 20 repetitions of CV	1,106 SNPs: 0.801	
					482 SNPs: 0.842	
[16]	ADNI	8,239 SNPs $ 471 subjects	SVM with mRMR filter, LASSO, RF, BSWiMS, RPART, KNN	CV	2,500 SNPs: 0.719	
					1000 SNPs: 0.554	
[58]	Genotype tissue expression & GWAS	57,853 Single nucleotide polymorphisms (SNPs)	RF	Dataset:90 % Training, 10% testing/ 10-fold CV	5785 SNPs(features): 0.75	70.63
					39 SNPs(features): 0.826	75.22

					11 SNPs(features): 0.8949	81.21
[50]		walking (gait) & cognitive tests data	SVM	Dataset:80 % Training, 20% testing/ 5-fold CV		78
[49]		Audio(speech) data	DT, bagging, MLP, LinearSVC, LRCV	Split dataset to multiple sets and use K-fold CV method		VBSD dataset: 83.3 Dem@Care dataset: 84.4

Table 6. A summary of Deep Learning approaches.

Ref No	Dataset/s	Data type	Algorithms	Evaluation technique	Best performance metrics	
					Area Under the Curve (AUC)	Accuracy (ACC)
[41]	OASIS	15,200 MRI of 170 AD & 70 NC	2D CNN	75% Training, 25% validation		99.3
[59]	ADNI	sMRI of 326 AD & 607 NC	lightweight 3D CNN	5-Fold Cross-Validation	0.9789	92.52
[60]	ADNI	amyloid PET	3D CNN	5-Fold Cross-Validation		90.8
[61]	ADNI	amyloid PET + MRI	2 identical CNN	5-Fold Cross-Validation		92.34
[62]	OptumLabs Data Warehouse (OLDW)	large amount of administrative health information	Feed Forward Network & Recurrent Neural Network (RNN)		MLP: 0.807 to 0.844 RNN: 0.77 to 0.843 RNN with pre-trained weights: 0.79 to 0.843	
[63]	ADNI	42,908,833 SNPs (from WGS) & 471 subjects	DNN & 1D CNN	5-Fold Cross-Validation	1000 SNPs: DNN= ≈ 0.55, CNN= ≈ 0.56 500 SNPs: DNN= ≈ 0.58, CNN= 0.57 200 SNPs: DNN= ≈ 0.62, CNN= ≈ 0.56	

					100 SNPs: DNN= ≈ 0.59, CNN= ≈ 0.55	
[17]	ADNI	486 SNPs + structural features of 56 brain regions (ROIs) & 632 subjects	3-layer convolutional neural network (CNN)	Dataset:80 % Training, 20% testing (5-fold CV)	Top 10: SNPs only= 0.65, SNPs +ROIs = 0.80 Top 20: SNPs only= 0.68, SNPs +ROIs = 0.73 Top 50: SNPs only= 0.68, SNPs +ROIs = 0.60 Top 100: SNPs only= 0.60, SNPs +ROIs = 0.60	
[18]	ADNI	41 SNPs (from GWAS) + 100 MRI scans	2D CNN & Multilayer perceptron network (MLP)	CNN: 5-fold CV for 100 scans, MLP: 300 SNPs train & validate, 75 SNPs test	SNPs only for 75 samples = 0.6807 MRI only for 100 samples = 0.8763 Refined CNN with SNPs for 100 samples= 0.8831	
[19]	ADNI	19 SNPs + APOE ε4 allele & MRI scans of 138 AD, 225 NC & 358 MCI	DNN	Dataset:80 % Training, 20% testing	SNPs only: 0.689 MRI only: 0.820 SNPs + MRI: 0.835	
[64]	GSE33000 and GSE44770 DS for gene expression data & GSE80970 DS for DNA methylation data	19,488 genes & 485,577 probes of CpG islands	DNN	5-Fold Cross-Validation	0.797	82.3

CONCLUSION

This survey explored some of recent approaches employing machine learning and deep learning algorithms to early predict Alzheimer's disease (AD) and contribute to its therapeutic development. These approaches were categorized in terms of learning technique and data modality used. In addition, they were discussed from different aspects, and their strengths, limitations and outcomes were compared. In spite of the great diversity of these approaches, almost all of them have endeavoured to offer the best model that could efficiently employ the medical dataset and successfully diagnose the disease. Nevertheless, some types of biomarkers such as genetic biomarkers were largely variant and complex. Therefore, this kind of data could dictate the type of algorithms used and the complexity level of the proposed model. It was noticed that most deep learning (DL) based approaches using genetic variants data tended to merge them with other modalities to improve the prediction accuracy, and this combination increased their complexity. On the other side, the other DL based approaches that used only genetic variants data could not achieve higher accuracy. Improving the prediction accuracy for AD using deep learning techniques with genetic variants data is still challenging. In the near future, we will propose a

deep learning model for predicting AD using genetic variants data. We will offer a pre-processing pipeline that seeks to reduce the complexity of these data and improve the prediction precision. This survey can be a coherent and informative reference for many researchers without a solid background in the latest AI technologies used for AD early diagnosis.

REFERENCES

[1] "2020 Alzheimer's disease facts and figures," *Alzheimer's Dement.*, vol. 16, no. 3, pp. 391–460, 2020, doi: 10.1002/alz.12068.

[2] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans, "Alzheimer disease in the US (2010-2050) estimated using the 2010 census," *Neurology*, vol. 80, no. 19, pp. 1778–1783, 2013.

[3] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and future treatments for Alzheimer's disease," *Ther. Adv. Neurol. Disord.*, vol. 6, no. 1, pp. 19–33, 2013, doi: 10.1177/1756285612461679.

[4] P. Kishore, U. C. Kumari, M. N. V. S. S. Kumar, and T. Pavani, "Detection and analysis of Alzheimer's disease using various machine learning algorithms," *Mater. Today Proc.*, vol. 45, no. xxxx, pp. 1502–1508, 2021, doi: 10.1016/j.matpr.2020.07.645.

[5] D. E. Barnes, and S. J. Lee, "Predicting Alzheimer's risk: Why and how?," *Alzheimer's Res. Ther.*, vol. 3, no. 6, pp. 1–3, Nov. 2011, doi: 10.1186/alzrt95.

[6] A. P. Porsteinsson, R. S. Isaacson, S. Knox, M. N. Sabbagh, and I. Rubino, "Diagnosis of Early Alzheimer's Disease: Clinical Practice in 2021," *J. Prev. Alzheimer's Dis.*, 2021, doi: 10.14283/jpad.2021.23.

[7] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," *Int. Conf. ICT Knowl. Eng.*, pp. 1–6, 2018, doi: 10.1109/ICTKE.2017.8259629.

[8] M. I. Jordan, and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (80-.).*, vol. 349, no. 6245, 2015.

[9] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016, doi: 10.1186/s13634-016-0355-x.

[10] N. K. Chauhan, and K. Singh, "A review on conventional machine learning vs deep learning," *2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018*, pp. 347–352, 2019, doi: 10.1109/GUCON.2018.8675097.

[11] S. Yang, J. M. S. Bornot, K. Wong-Lin, and G. Prasad, "M/EEG-Based Bio-Markers to Predict the MCI and Alzheimer's Disease: A Review from the ML Perspective," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2924–2935, 2019, doi: 10.1109/TBME.2019.2898871.

[12] X. Wang, J. Qi, Y. Yang, and P. Yang, "A survey of disease progression modeling techniques for alzheimer's diseases," *IEEE Int. Conf. Ind. Informatics*, vol. 2019-July, pp. 1237–1242, 2019, doi: 10.1109/INDIN41052.2019.8972091.

[13] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," *2020 Int. Conf. Emerg. Technol. INCET 2020*, no. December, 2020, doi: 10.1109/INCET49848.2020.9154130.

[14] T. Jo, K. Nho, and A. J. Saykin, "Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data," *Frontiers in Aging Neuroscience*, vol. 11. 2019, doi: 10.3389/fnagi.2019.00220.

[15] E. Lin, C.-H. Lin, and H.-Y. Lane, "Deep Learning with Neuroimaging and Genomics in Alzheimer's Disease," *Int. J. Mol. Sci.*, vol. 22, no. 15, p. 7911, 2021, doi: 10.3390/ijms22157911.

[16] J. De Velasco Oriol, E. E. Vallejo, K. Estrada, J. G. Taméz Peña, and T. A. s. Disease Neuroimaging Initiative, "Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data," *BMC Bioinformatics*, vol. 20, no. 1, p. 709, Dec. 2019, doi: 10.1186/s12859-019-3158-x.

[17] J. Zhou, L. Hu, Y. Jiang, and L. Liu, "A Correlation Analysis between SNPs and ROIs of Alzheimer's Disease Based on Deep Learning," *Biomed Res. Int.*, vol. 2021, 2021, doi: 10.1155/2021/8890513.

[18] Q. Ying, X. Xing, G. Liang, and I. City, "Multi-Modal Data Analysis for Alzheimer ' s Disease Diagnosis : An Ensemble Model Using Imagery and Genetic Features," *bioRxiv*, pp. 5–10, 2021.

[19] K. Ning *et al.*, "Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework," *Neurobiol. Aging*, vol. 68, pp. 151–158, 2018, doi: 10.1016/j.neurobiolaging.2018.04.009.

[20] J. Williamson, J. Goldman, and K. S. Marder, "Genetic aspects of alzheimer disease," *Neurologist*, vol. 15, no. 2. pp. 80–86, Mar. 2009, doi: 10.1097/NRL.0b013e318187e76b.

[21] Q. Sun, N. Xie, B. Tang, R. Li, and Y. Shen, "Alzheimer's disease: From genetic variants to the distinct pathological mechanisms," *Front. Mol. Neurosci.*, vol. 10, no. October, pp. 1–14, 2017, doi: 10.3389/fnmol.2017.00319.

[22] R. Mishra, and B. Li, "The application of artificial intelligence in the genetic study of Alzheimer's disease," *Aging Dis.*, vol. 11, no. 6, pp. 1567–1584, 2020, doi: 10.14336/AD.2020.0312.

[23] A. T. Isik, "Late onset Alzheimer's disease in older people.," *Clin. Interv. Aging*, vol. 5, pp. 307–311, 2010, doi: 10.2147/cia.s11718.

[24] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Arch. Dis. Child. Educ. Pract. Ed.*, vol. 98, no. 6, pp. 236–238, Dec. 2013, doi: 10.1136/archdischild-2013-304340.

[25] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," 2006. doi: 10.1016/j.ajog.2006.07.001.

[26] G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati, and E. Abadie, "Genetic tests and genomic biomarkers:

Regulation, qualification and validation," 2008.
[27] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine learning SNP based prediction for precision medicine," *Frontiers in Genetics*, vol. 10, no. MAR. Frontiers Media S.A., 2019, doi: 10.3389/fgene.2019.00267.
[28] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.
[29] R. Tarawneh, "Biomarkers: Our Path Towards a Cure for Alzheimer Disease," *Biomark. Insights*, vol. 15, 2020, doi: 10.1177/1177271920976367.
[30] S. Lovestone, "Biomarkers in Alzheimer's disease," *Res. Pract. Alzheimers. Dis.*, vol. 11, no. 1, pp. 41–50, Mar. 2006, doi: 10.1515/almed-2020-0090.
[31] W. M. van Oostveen and E. C. M. de Lange, "Imaging techniques in alzheimer's disease: A review of applications in early diagnosis and longitudinal monitoring," *Int. J. Mol. Sci.*, vol. 22, no. 4, pp. 1–34, 2021, doi: 10.3390/ijms22042110.
[32] G. Gifford, R. McCutcheon, and P. McGuire, "Neuroimaging studies in people at clinical high risk for psychosis," *Risk Factors Psychos.*, pp. 167–182, Jan. 2020, doi: 10.1016/b978-0-12-813201-2.00009-0.
[33] S. Al-Shoukry, T. H. Rassem, and N. M. Makbol, "Alzheimer's diseases detection by using deep learning algorithms: A mini-review," *IEEE Access*, vol. 8, pp. 77131–77141, 2020, doi: 10.1109/ACCESS.2020.2989396.
[34] G. S. Bloom, "Amyloid-β and tau: The trigger and bullet in Alzheimer disease pathogenesis," *JAMA Neurol.*, vol. 71, no. 4, pp. 505–508, 2014, doi: 10.1001/jamaneurol.2013.5847.
[35] P. A. Rowley, A. A. Samsonov, T. J. Betthauser, A. Pirasteh, S. C. Johnson, and L. B. Eisenmenger, "Amyloid and Tau PET Imaging of Alzheimer Disease and Other Neurodegenerative Conditions," *Semin. Ultrasound, CT MRI*, vol. 41, no. 6, pp. 572–583, Dec. 2020, doi: 10.1053/j.sult.2020.08.011.
[36] A. J. Saykin *et al.*, "Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans," *Alzheimer's Dement.*, vol. 6, no. 3, pp. 265–273, 2010, doi: 10.1016/j.jalz.2010.03.013.
[37] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007, doi: 10.1162/jocn.2007.19.9.1498.
[38] J. Tian *et al.*, "Modular machine learning for Alzheimer's disease classification from retinal vasculature," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-020-80312-2.
[39] P. J. Snyder *et al.*, "Retinal imaging in Alzheimer's and neurodegenerative diseases," *Alzheimer's Dement.*, vol. 17, no. 1, pp. 103–111, 2021, doi: 10.1002/alz.12179.
[40] C. Sudlow *et al.*, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLoS Med.*, vol. 12, no. 3, Mar. 2015, doi: 10.1371/journal.pmed.1001779.
[41] F. E. K. Al-Khuzaie, O. Bayat, and A. D. Duru, "Diagnosis of Alzheimer Disease Using 2D MRI Slices by Convolutional Neural Network," *Appl. Bionics Biomech.*, vol. 2021, 2021, doi: 10.1155/2021/6690539.
[42] M. L. Metzker, "Sequencing technologies the next generation," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010, doi: 10.1038/nrg2626.
[43] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012, doi: 10.1371/journal.pcbi.1002822.
[44] C. S. Ku, E. Y. Loy, A. Salim, Y. Pawitan, and K. S. Chia, "The discovery of human genetic variations and their use as disease markers: Past, present and future," *J. Hum. Genet.*, vol. 55, no. 7, pp. 403–415, 2010, doi: 10.1038/jhg.2010.55.
[45] A. Anjum, S. Jaggi, E. Varghese, S. Lall, A. Bhowmik, and A. Rai, "Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach," *J. Comput. Biol.*, vol. 23, no. 4, pp. 239–247, 2016, doi: 10.1089/cmb.2015.0205.
[46] S. Perera, K. Hewage, C. Gunarathne, R. Navarathna, D. Herath, and R. G. Ragel, "Detection of Novel Biomarker Genes of Alzheimer's Disease Using Gene Expression Data," *MERCon 2020 - 6th Int. Multidiscip. Moratuwa Eng. Res. Conf. Proc.*, pp. 1–6, 2020, doi: 10.1109/MERCon50084.2020.9185336.
[47] J. F. Loring, X. Wen, J. M. Lee, J. Seilhamer, and R. Somogyi, "A gene expression profile of Alzheimer's disease," *DNA Cell Biol.*, vol. 20, no. 11, pp. 683–695, 2001, doi: 10.1089/10445490152717541.
[48] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011, doi: 10.1093/bioinformatics/btr330.
[49] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying Alzheimer's disease," *Simul. Model. Pract. Theory*, vol. 99, p. 102023, 2020, doi: 10.1016/j.simpat.2019.102023.
[50] B. Ghoraani, L. N. Boettcher, M. D. Hssayeni, A. Rosenfeld, M. I. Tolea, and J. E. Galvin, "Detection of mild cognitive impairment and Alzheimer's disease using dual-task gait assessments and machine learning," *Biomed. Signal Process. Control*, vol. 64, no. August 2020, p. 102249, 2021, doi: 10.1016/j.bspc.2020.102249.
[51] A. Khan, and S. Zubair, "An Improved Multi-Modal based Machine Learning Approach for the Prognosis of Alzheimer's disease," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2020.04.004.
[52] Y. Zheng, H. Guo, L. Zhang, J. Wu, Q. Li, and F. Lv, "Machine learning-based framework for differential

diagnosis between vascular dementia and Alzheimer's disease using structural mri features," *Front. Neurol.*, vol. 10, no. OCT, pp. 1–9, 2019, doi: 10.3389/fneur.2019.01097.

[53] V. P. S. Rallabandi, K. Tulpule, and M. Gattu, "Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis," *Informatics Med. Unlocked*, vol. 18, 2020, doi: 10.1016/j.imu.2020.100305.

[54] J. H. Park *et al.*, "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," *npj Digit. Med.*, vol. 3, no. 1, 2020, doi: 10.1038/s41746-020-0256-0.

[55] M. Grassi *et al.*, "A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures," *Front. Neurol.*, vol. 10, no. JUL, pp. 1–15, 2019, doi: 10.3389/fneur.2019.00756.

[56] N. Arzouni, W. Matloff, L. Zhao, K. Ning, and A. W. Toga, "Identification of Dysregulated Genes for Late-Onset Alzheimer's Disease Using Gene Expression Data in Brain.," *J. Alzheimer's Dis. Park.*, vol. 10, no. 6, 2020, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33282526%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7717689.

[57] 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, pp. 1–15, 2020, doi: 10.1371/journal.pone.0232103.

[58] U. Rangaswamy, S. A. P. Dharshini, D. Yesudhas, and M. M. Gromiha, "VEPAD - Predicting the effect of variants associated with Alzheimer's disease using machine learning," *Comput. Biol. Med.*, vol. 124, no. August, p. 103933, 2020, doi: 10.1016/j.compbiomed.2020.103933.

[59] S. Katabathula, Q. Wang, and R. Xu, "Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations," *Alzheimer's Res. Ther.*, vol. 13, no. 1, pp. 1–9, 2021, doi: 10.1186/s13195-021-00837-0.

[60] T. Jo, K. Nho, S. L. Risacher, and A. J. Saykin, "Deep learning detection of informative features in tau PET for Alzheimer's disease classification," *BMC Bioinformatics*, vol. 21, no. 21, pp. 1–14, 2020, doi: 10.1186/s12859-020-03848-0.

[61] A. Punjabi, A. Martersteck, Y. Wang, T. B. Parrish, and A. K. Katsaggelos, "Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks," *PLoS One*, vol. 14, no. 12, pp. 1–14, 2019, doi: 10.1371/journal.pone.0225759.

[62] V. S. Nori, C. A. Hane, Y. Sun, W. H. Crown, and P. A. Bleicher, "Deep neural network models for identifying incident dementia using claims and EHR datasets," *PLoS One*, vol. 15, no. 9 September, pp. 1–12, 2020, doi: 10.1371/journal.pone.0236400.

[63] J. de Velasco Oriol, E. Vallejo, and K. Estrada, "Predicting late-onset Alzheimer's disease from genomic data using deep neural networks," *bioRxiv*, p. 629402, 2019, doi: 10.1101/629402.

[64] W. J. Lim, K. H. Kim, J. Y. Kim, S. Jeong, and N. Kim, "Identification of DNA-methylated CpG islands associated with gene silencing in the adult body tissues of the ogye chicken using RNA-Seq and reduced representation bisulfite sequencing," *Front. Genet.*, vol. 10, no. APR, p. 346, 2019, doi: 10.3389/fgene.2019.00346.

[65] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Syst. Appl.*, vol. 140, p. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.

[66] S. Mishra, A. Dash, and L. Jena, "Use of deep learning for disease detection and diagnosis," in *Studies in Computational Intelligence*, vol. 903, Springer, 2021, pp. 181–201.

[67] L. Koumakis, "Deep learning models in genomics; are we there yet?," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1466–1473, 2020, doi: 10.1016/j.csbj.2020.06.017.