Early Alzheimer's Disease Detection Using Different Techniques Based on Microarray Data: A Review

https://doi.org/10.3991/ijoe.v18i04.27133

Shaymaa Taha Ahmed[1,2](✉), Suhad Malallah Kadhem[1]
[1] University of Technology, Baghdad, Iraq
[2] University of Diyala, Diyala, Iraq
Shaymaa.taha.ahmed@basicedu.uodiyala.edu, mrs.sh.ta.ah@gmail.com

Abstract—Alzheimer's Disease (AD) is a degenerative disease of the brain that results in memory loss due to the death of brain cells. Alzheimer's disease is more common as people get older. Memory loss happens over time, and as a result, the person loses the ability to react appropriately to their surroundings. Microarray technology has emerged as a new trend in genetic research, with many researchers utilizing it to look at the changes in gene expression in particular organisms. Microarray experiments can be used in various ways in the medical field, including the prediction and detection of disease. Large amounts of unprocessed raw gene expression profiles sometimes contribute to computational and analytic difficulties, including selecting dataset features and classifying them into an appropriate group or class. The large dimensions, lesser sample size, and noise in gene expression data make it difficult to attain good Alzheimer classification accuracy using the entire collection of genes. The categorization process necessitates careful feature reduction. As a result, a comprehensive review of microarray Alzheimer's disease studies is presented in this paper, focusing on feature selection techniques.

Keywords—feature selection, Alzheimer's disease, microarray technology, machine learning, deep learning

1 Introduction

Over time, memory loss and intellectual impairment occur as a result of AD, a common form of dementia, as well as other mental functions being impaired. AD causes structural alterations in the brain. The symptoms appear gradually and steadily worsen with time. The patient first develops Moderate Cognitive Impairment (MCI), which then advances for Alzheimer's. AD, Moderate Cognitive Impairment stage occurs halfway through the disease. Though not all MCI patients develop AD [1][2],some do. Although AD is currently incurable, it can be slowed or stopped in its tracks if caught early [3].In 2006, 26.6 million persons were diagnosed with AD. In 2050, According to current estimates, AD will affect 1 in every 85 people sometime in the future, with around 43% of cases requiring high-level care [4].The transition from a healthy state to AD

can take many years in Alzheimer's patients [5].Patients first acquire Moderate Cognitive Impairment, which then progresses to Alzheimer's Disease. However, not all MCI patients get AD [6].As a result, the current study is mostly focused on predicting the change of MCI to AD.

In the last decade, Alzheimer's Disease diagnosis has benefited greatly from the use of machine learning techniques [7][8]. Support vector machines are the most extensively used methods of classification, (ANNs), also Deep Learning (DL). This nature of a problem of optimization [9][10][11]is the major difference between SVM and ANN. SVM provides a globally optimal solution. [12][13][14]ANN, on the other hand, provides a locally optimal solution. Feature extraction is a key stage in both SVM and ANN. [15][16]suggested that combining neural networks with intelligent agents to medical image processing could be advantageous. Deep learning, on the other hand, includes the feature extraction procedure directly into the learning model [17][18][19].When dealing with huge datasets, Deep learning has been proven to be particularly beneficial for picture data [17][20].Some researchers employed ensemble approaches to increase Alzheimer's disease classification accuracy [21][22][23].

Deep exploration of molecular pathways has been more common in recent decades as a strategy of research for discovery in effect treatments used for complicated illnesses such as Alzheimer’s, cancer, diabetes, as well as other diseases. Microarrays and Next-Generation Sequencing(NGS) are most often used technology in research methods. Studying using in-depth inquiry approaches have focused on the processing stages, it consists of a variety of feature selection(FS) as well as dimension reduction techniques [24][25]. There are two types of people who work in this sector. The initial group of investigations [26][27]used image recognition methods due to brain scan data (such as MRI). The research in the second group, used data from gene expression(GE) to estimate the risk of developing Alzheimer's illness [28][29].High dimensionality data were found in gene expression data, however, including irrelevant, the disease diagnosis was unaffected by the redundant and noisy genes. Medical diagnostic accuracy is hampered by the application of artificial intelligence and data mining techniques because the prevalence for redundancy data expression and the limited sample size versus the huge number of genes (features) [30][31]. As a result, dimensionality reduction is an exciting topic of data mining, statistics pattern recognition, and machine learning.

Dimensionality reduction (DM) aims to increase a classification algorithm's accuracy by reducing redundant and meaningless data from the Microarray dataset. There are numerous methods for reducing the dimensionality of a system. Dimensionality reduction methods are determined by the application domain and the dataset's peculiarities. Filter, wrapper, embedding, and hybrid approaches are all types of feature selection strategies [32]. Filter algorithms select attributes based on the characteristics of particular users [33][34]. Wrapper techniques make use utilizing machine learning techniques or population movements; a subset of features is selected.

Methods that use filters are well-known for their ability to do calculations quickly at the cost of accuracy, whereas wrapper approaches have superior accuracy performance while requiring less computation. In domains with large datasets, filter-based method proved to be faster than wrapper methods. Both approaches have a drawback in that they fail to take into account how the classifier interacts with the addictions between

the different features, resulting in varying classification accuracy depending on which features are used. Embedded techniques, on hand, Use learning algorithms to improve your performance. Wrapper approaches take a higher processing cost, while embedding methods offer the benefit of interfacing the with classification system [32].

Many methods of feature selection with in research are aimed on selecting useful and relevant traits that increase classification rates while decreasing computing costs, limited of these studying, however, in fact look at all of the methods used. The large majority of evaluations focused on a specific strategy for selecting features [35][36][37]. The medical industry's general use of feature selection [38]. All available methodologies with their taxonomy in earlier research were not offered a thorough state-of-the-art, and issues with microarray data and experiments were not addressed. As an illustration, consider [36]. gave a review of filtering strategies for microarray gene analysis [37]. however, microarray data was not the focus of the survey on feature selection and fusion algorithms at the feature level. Gene expression microarray data were selected using matched-pairs feature selection, which was provided in a compressively brief overview by [39]. [40]gave a quick overview of widely used feature extraction and feature selection approaches that are currently popular. An overview of feature selection methodologies for medical field challenges was also provided by [41]. which incorporates biological, medical imaging, signal processing, and DNA microarray analysis of data. As a result, the study published by [42] is the most nearly similar to this review; nevertheless, Feature selection was all that was discussed in their review, the sources of microarray datasets and the problems they raise.

The objectives of this paper is to provide information on the difficulties and problems associated with microarray Alzheimer's datasets, with the current the feature selection methods employed in feature selection, Describe the microarray experiment in detail and point out the limitations of current methods. This publication also outlines important research opportunities for the future in this field.

All of this following information may be found in the paper: Section 2 provides an overview of Microarray technique and accompanying data, Section 3 focuses on Gene Selection. Dimension Reduction Approaches are classified in Section 4 according to their taxonomy. In Section 5, open research concerns are discussed. These concerns were taken into account when the data was collected. Section 6 discusses the literature review; Section 7 provides as the paper's conclusion.

2 Microarray technique

Microarray Technology (MT) is making revolutionary advances in the biological sciences since its introduction. It's seen as a launching pad for important new studies. It's made it look at tens of thousands of gene activity at once. even if most biologists and other researchers have problems when mining and working with for this data type. There are various databases where the results of Microarray experiments can be found.

The use of microarrays in scientific research dates back to the mid-1980s [43]. DNA Microarrays were first described by Augenlicht et al. (1987), who discovered over 4000

complementary DNA (cDNA) sequences on nitrocellulose [44]. Microarrays have allowed biologists to look into and measure for expression of tens of thousands of genomes at the same time [43][38]. Bioinformatics, medical areas, all of these fields, as well as microarray research, have profited from advances in the technology [40].This type of microarray is often referred to as a biochip or maybe a DNA chip it has the number from little patches of genetic material attached a solid surface. DNA Microarrays are being used by scientists as a stage for studying the points of expression from different gene at same time, the numerous components of a person's genotype [43].

2.1 Microarray data

It is common practice to arrange and save Microarray experiment data in big matrices (M N). In reference to Table 1, There are rows of samples and columns of genes in every Microarray data matrix (features).

M by N matrices, which contain microarray data, are very huge. where N is the numbers of column and M is number of rows, each cell in a sample has its own unique value for gene expression [45][46]. X ij shows the levels of expression of genes j then the situation or else sample i. while j is a positive integer between 1 and M, and I is a negative integer from 1 to N.

Table 1. A matrix of microarray data

samples	**Gene Expression**		
	G1	G2 ----------	GM
S1	Y11	Y12	Y1M
S2	Y21	Y22	Y2M
-			
-			
SN	YN1	YN2	YNM

2.2 Microarray analysis of data

Because of the vast amount of data that can be extracted from genetic tissues, In the recent few decades, the biomedical industry has been increasingly important for machine learning. DNA Microarray datasets, in particular, have paved the way for the establishment of a new and dynamic area of bioinformatics research machine learning and. When there are so few samples (generally less than 100), yet so many properties, microarray data are usually viewed a structured data for machine learning (in the order of thousands).

Research groups in machine learning face significant obstacles when dealing with data of this type, since "false positives" are a possibility, or when selecting relevant features the prediction model, problems can arise (genes) [38]. Few genes a DNA Microarray are useful classification, according to research published in the literature. The removal of redundant and irrelevant information is critical in this case, as well as assisting experts in uncovering basic links between gene expression and certain diseases.

As a result, the gene dataset must be reduced in order to lower the expense of finding the optimal genes for differentiating between cell types (normal or abnormal cells). Clustering and classification [47] are well-known approaches for in-depth Microarray data analysis. Classification by clustering is an unsupervised strategy for sorting large amounts of data mad about smaller collections of genes otherwise samples that take common traits or shapes. Classes are created by the use of examples in a supervised learning environment. The classifier studies to classify unidentified sample instances into single of the specified sessions when given a batch of reclassified samples [48].

3 Background and advance of Gene Selection

To remove genes from a gene expression data, you can use a technique called Gene Selection, the DNA microarray is an example of this, that are redundant and/or ineffective. Feature selection is based on machine learning to select genes, which is well-suited for applications involving thousands of characteristics [49][50]. Using Gene Selection approaches, scientists aim to locate and eliminate duplicate genes in the original space in two ways: first, to find and express the most useful information. In theory, Overfitting will reduce generalization and degrade model performance if the number of genes is increased. Our current work for Gene Selection (GS) is primarily concerned with identifying the most important genes, with less attention paid to reducing irrelevant or redundant genome [51]. Relevance, redundancy, and complementarity must be prioritized if meaningful outcomes are to be achieved. The relevance of a gene is determined by whether or not it possesses the appropriate info about the given class. According to [52], the feature set can be divided into three categories: highly relevant, marginally relevant, and irrelevant. There two types of marginally relevant features: those that are redundant and those that are not The non-redundant and highly relevant feature sections contain the vast majority of the useful material [53].Genes are chosen using a similar set of algorithms based on microarray data (Figure 1).

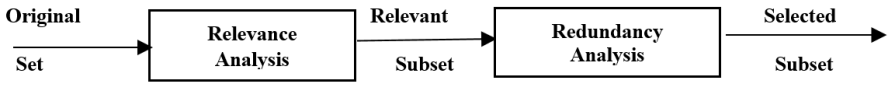

Fig. 1. A framework of Gene Selection

4 Dimension reduction methods

When dealing with large amounts of dimensional data, classification algorithms face numerous computational and memory challenges [5]. There are two approaches to reduce the dimensionality of a system: extraction and selection of characteristics (also called Dimensionality Reduction (DR)/ Feature Transformation(FT)). There is no indication about the importance for a sole feature missing while using the Feature Selection method, the only exception being when multiple unique features are required, this may lead to the loss of information if certain features are missed when selecting a feature

subset. Feature extraction, in contrast, the feature set can be reduced without losing too much information about the original feature. A type of data and application domain effect the selection of feature extraction and feature selection approaches.

4.1 Feature selection

High dimensional dataset includes features that are redundant, deceptive, or both, making it more difficult to interpret data further and thus not adding to process of learning. It's called feature subset selection when you pick the greatest characteristics from all the ones that can be utilized to differentiate between classes. a certain definition of relevance activates the statistical method called the feature selection algorithm. [54][55] Many feature selection methods have been empirically evaluated. According to various evaluation criteria, the search problem is a common term for feature selection. Feature selection algorithms use a searching organization that's characterized by exponential, sequential, or random search methods are all potential types of searching. It is possible to explore five various operators to create successors; weighted, compound, and random are a few of the possibilities available. Evaluation Metrics: Probability of Error, Divergence, Dependence, and Interclass Distance can be used to evaluate successors, Evaluation of Information, Uncertainty, and Stability in Figure 2 shows.

Filters, wrappers, and embedded/hybrid techniques are the three main types of feature selection methods. Wrapper-based approaches improve methods that use filters because of Feature Selection (FS) process is specific to a classifier being employed. Nevertheless, Wrapper techniques are very expensive to use in large feature spaces because to their high processing costs, and a trained classifier must be used to evaluate each feature set, this makes the process of selecting features more time consuming. In comparison to wrapper approaches, Filtering techniques are more efficient and faster to compute than traditional methods, however, their classification reliability is inefficient, making them better suited for large, complex datasets. Ways combining hybrid and embedded, which combine the best features of filters and wrappers, have recently been created. Combining independent tests with performance assessment functions for a feature subset is a hybrid approach [56][57]. There are two classes of filter techniques, Specifically, feature weighting techniques and subset selection methods are shown in Figure 2. Methods for weighing in consider each feature separately and assign a value to it based on how important it is to the overall aim [58].

Following advantages of feature selection have been made available to you:

- It decreases the feature space's dimensionality, hence reducing storage requirements and speeding up algorithms.
- Data that is redundant, irrelevant, or obtrusive is discarded using this method.
- Speeding up the learning algorithms' execution time has direct effects on data analysis activities.
- Enhancing the accuracy of the data.
- Increasing the resulting model's accuracy.
- Reduction of the feature set in order to conserve resources for the next round of data gathering or during usage.

- Enhancement of capabilities in order to increase prediction accuracy.
- Understanding data to learn more about the process that generated it or simply to see how it looks.

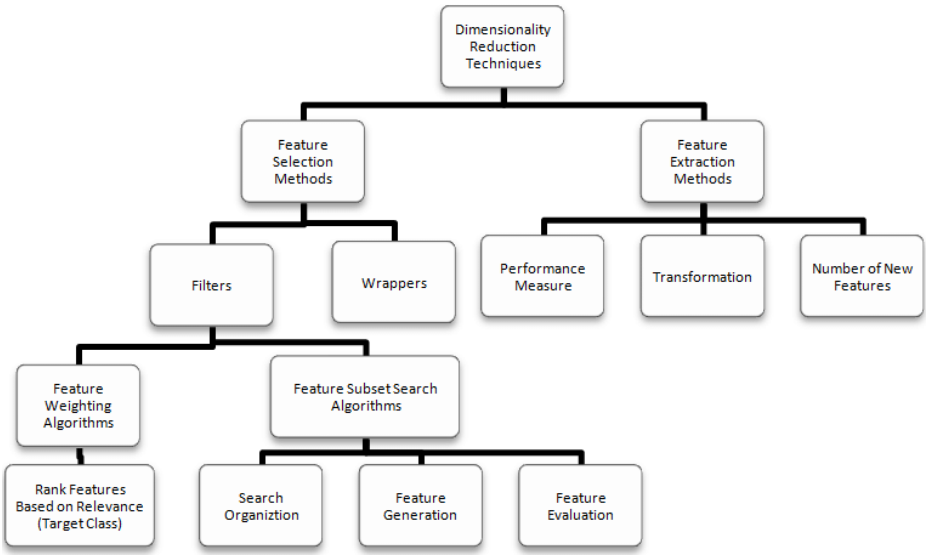

Fig. 2. Dimensionality reduction approaches: A hierarchical structure

4.2 Feature transforming extracting

Extracting features is a process that involves transforming the original features into more significant ones. extracted features according to the following definitions: "Feature extraction" primarily refers to the process of creating linear combinations αTx of continuous features with high discrimination power between groups. Finding a good representation for multivariate data is a significant topic in artificial intelligence (AI) and neural networks (NN). In this case, features extraction can be used to simplify the data and offer a linear combination of every variable in feature set with the original input variable as input [59][60]. Is the most extensively utilized feature extraction method. PCA a plethora of variations that have been put out. Principal component analysis is now an easy-to-understand analysis, the most important information in a confusing and redundant data collection can be found using a non-parametric methodology. Using principal component analysis (PCA) , we may reduce duplication (measured by covariance) while increasing info (measured by variance) in our data [61][62] [63] .

Many alternative dimensionality reduction strategies had really been established examined on two separate kinds of data to see how they affect classification performance, including information gain, wrapper approaches, or feature extraction through a variety of PCA techniques (e-mail data , medicine detection data) [64][65].There is a strong

correlation between the type of data and the results of PCA feature extraction (transformation). For both types of data, the method of deciding which features to include Wrapper has a moderate impact on categorization accuracy as compared to information gain.

As a result of a research, it's clear that dimensionality reduction is critical. Comparing feature selection methods with feature extraction methods, wrappers produce smaller feature subsets with better classification accuracy. Though computationally more expensive than feature extraction algorithms, wrappers are a viable alternative [66][67]. in [56][68] proposed methods for lowering the dimensionality of feature extraction and feature selection on the bi-level, as a means to enhance categorization efficiency. Dimensionality reduction begins with this step is to choose features depending on how closely they relate to one another. Select features from first stage are used in PCA and LPP at the second level to extract additional features. The suggested method was tested on a variety of widely used datasets to see how well it worked. The findings obtained suggest that the proposed system outperforms single-level dimensionality reduction approaches.

5 Literature review

In the literature, several approaches to Alzheimer's disease are employed in numerous pieces of work (AD). This section will include illustrations of the most recent research done in the field.

In 2011, B. Booij ,et al[69]. A disease classifier algorithm was developed using a Jackknife gene selection(GS) technique and (PLSR), which provides a test score indicating whether Alzheimer's disease (AD) is present or not (negative). An independent test group of 63 people, including 31 (AD) patients, 25 (HC) of the same age, and 7 young controls, validated the algorithm, which relies on 1239 probes. This technique accurately predicted of 55/63. (AUC 87 %).

In 2012, L. Scheubert, et al [70]. selection of features utilizing three different methods: (IG), (RF) accuracy, (GA) and Support Vector Machine (SVM) wrapper. When evaluating their output, we contrast it with GA/SVM outcomes (accuracy 85 percent). For the reason that of the lesser sample sizes in addition to unstable nature of this algorithm being presented.

In 2013,k, Lunnon, et al [71]. T-tests utilizing Meng scores and backward are two approaches for testing hypotheses that have been presented. we acquired a 75% accuracy rate in the validation group using AD and a control device. Sample sizes are restricted since they are small.

In 2014, P,Johnson, et al [72]. In this paper used Genetic algorithms (GA), as in the prediction of the onset of AD. An Accuracy of 0.90 for predicting HC and 0.86 for MCI conversion at (36) months that has been cross-validated. The constraints of the paper are as follows, the model developed is difficult to decipher, and the available data is less prone to overfitting.

In 2015, F, Sherif, et al [73]. The efficiency of the Bayesian network (BN) in determining the causes of SNPs has been demonstrated with a respectable level of precision. A result or included with indicated for advantage of a SNP group found using during

this Markov techniques, does have a strong connection to AD and outperforms both the Nave Bayes(NB), the nave tree fed Bayes(NTB). This idea on building medicinal techniques for drug discovery is still completed. The accuracy and sensitivity of the minimal enhanced Markov blanket are 66.13 percent and 88.87 percent, respectively, compared to 61.58 percent and 59.43 percent in naive Bayes.

In 2015 ,S, Sood, et al [74]. To predict HC conversion to MCI/AD, we used Bayesian statistics (ULSAM Ageing) and KNN with AUC of 0.73%. In most cases, the microarray data are three-dimensional or more. sample sizes and variables that are not important to the study are covered in large numbers. Generate a lot of noise. As a result, finding out about the data sets and looking for correlations between qualities might be challenging.

In 2015,S, Paylakhi, et al [75]. The (GA) and (SVM) have been employed to build a gene selection strategy in this study. To begin, Using Fisher criteria, High dimensional microarray data could have noise and redundant gene eliminated. A (GA-SVM) then using to choose distinct subsets of maximally informative genes using different training sets. The Fisher Score and (GA)(SVM) approaches that combined for profit of a filtering technique and combined way. The suggested technique was evaluated using (AD) DNA microarray data. The result shows the suggested technique has a strong performance in classification and selection, which may provide a classification accuracy of 100 percent with only 15 genes. restrictions due to the detail that gene expression (GE) data can been erroneous or else missing.

In 2016 , S,Zahra Paylakhi, et al [76]. These methods combine the fisher Score, significant analysis of microarrays, and a (GA)- (SVM). A Fisher technique is employed for remove redundant and noisy genes from microarray data. Genetic algorithm - (SVM) selects subsets of highly informative genes using different training sets and the SAM approach is usage. Microarray data from AD patients was usage for test the proposed technique. The result appearances that suggested method implements fit in selection with classification, It has a classification accuracy of 94.55% utilizing just 44 genetic parameters. Biologically speaking, at least 24 (55%) of these genes are related with dementia, namely Alzheimer's disease. Small sample sizes and low precision limit the ability to combine datasets from various sources in order to improve precision.

In 2016,N, Voyle, et al [77]. Methods: for predicted used random forest (RF) and removal of the recursion feature. All analyses included age and APOE 4 genotype as variables. 70 percent of the time. We discovered that a lack of homogeneity among the control group may have resulted in lower prediction accuracy.

In 2016,M, Barati, et al [78]. Methods include (SVM), information, deviation, Gini coefficient and the gain ratio. A minimum of two algorithm weights greater than 0.5 are considered important for the sequences studied. A neural network approach (such as auto multilayer perceptron, neural net, and perceptron) was then applied to 11 sets of data using the weighted perceptron technique, with an overall performance of 97 percent. It does, however, introduce some issues since even if features have been selected, they do not provide the same level of confidence as a stepwise selection process that goes in both directions.

In 2017, M, Balamurugan, et al [79]. They proposed KNN Classifying Algorithm according to dimensionality reduction for diagnosing and classification Alzheimer's

disease(AD), (MCI) in datasets. The (RDD-UDS) is a dataset provided by the (NACC) enabling researchers to analyze clinical and statistical dataset. The drawbacks of the KNN method based on the feature from a data; with huge data, the prediction step may be slow and sensitive to the data's size and irrelevant aspects.

In 2017,K, Nishiwaki, et al[80]. machine learning technology of random forest to develop a gene selection method. A study with an accuracy of 0.83 percent employed this method on (AD) microarray data to appropriately score the gene. The main weakness all datasets used are microarrays, hence their RNA-seq application is more accurate and less noisy.

In 2017,H, Li, et al [81]. proposed a method, The Ref-REO assay is used to identify variations in leukocyte-specific expression in blood samples containing both white and red blood cells. We found 42 and 45 DEGs in two datasets using Ref-REO in this work, which compared Alzheimer's disease (AD) blood samples to normal peripheral whole blood (PWB), with an AUC greater than 0.73 for predicting AD .It's quite tough to choose an appropriate feature combination from little DNA microarray data that's high dimensional.

In 2018, L, Xu, et al [82]. Alzheimer's disease should be detected at an early stage, scientists have developed a computational method analysis of protein sequence data. The number of times two amino acids appear in a row is used in their improved technique to represent sequences, and the SVM classifies the data after that. Magnetic resonance imaging-based research has been done in the past, but this new approach is more expensive and time demanding. Experiments have shown that the approach they designed has an accuracy of 85,7 percent. Additionally, the dataset used to classify AD their efforts resulted in the creation of. The main weakness in their system is that they don't look at how qualities interact with one another to improve predictions method.

In 2018,X, Li, et al [83]. In this paper, first big systematic analysis was done to discover (DEGs) had samples of blood with (245) Alzheimer's disease, 143 (MCI), and 182 (HC). A genome-wide association analysis was conducted to identify novel risk genes based on gene-based analyses of two different datasets of Alzheimer's disease blood samples. There was a new test that could tell Alzheimer's disease patients of healthy controls with a precision of 85.7 %. Limitation a small number of features.

In 2019, K, Sekaran, et al [84]. In this work, the gene expression profiles of Alzheimer's disease (AD) and healthy individuals are compared using numerical methods and (ML) techniques. Identification of differential gene expression) contributes significantly to the identification of most useful genes. Rhinoceros Search Technique, an algorithm based on a meta-heuristic globally optimization meta-heuristic (RSA). In the wake of RSA, researchers have discovered 24 new gene biomarkers. Four supervised ML techniques including Support Vector Machines, Random Forest , Nave Bayes and (MLP-NN) are usage to classify two separate groups of samples. One of these models, the RSA-MLP-NN, was 100 percent accurate in distinguishing between Alzheimer's disease (AD) and normal genes, demonstrating its usefulness. The study's weakness is that the training set is possible to contain a large amount from noise, which could have an impact on model performance.

In 2020,T, Lee, et al [85]. For the aim of this research. Five (5) feature selection approaches and five classifications have been used to identify genes related with Alzheimer's disease and to differentiate those patients. The best average AUC values for ADNI, ANMI, and ANM2 were 0.657, 0.874, and 0.804. For external validation, the greatest accuracy was 0.697 (for training ADNI to test ANM1) value 0.76 (for ADNI-ANM2) value 0.61 (for ANM1-ADNI) value 0.79 (for ADNI-ADN2), and 0.655 (for ANM2-ADNI), with an overall AUC of 0.859. (ANM2-ANM1). Due to sample size limits and low accuracy, a combination of feature selection approaches and local search methods was used to improve accuracy.

In 2020, H,Ahmed, et al [86] . The focus of this research is on the use of ML approaches to identify AD biomarkers. Random Forest (RF), Nave Bayes (NB), (LR) and Support Vector Machine algorithms were used to every Alzheimer's disease genetic information from ADNI-1 imaging project datasets. Nave Bayes (NB), Random Forest (RF), Support Vector Machine(SVM), and Logistic Regression methods got 98.1 percent, 97.97 percent, 95.88 percent, and 83 percent overall accuracy in ADNI-1's whole-genome approach. The findings suggest that classification algorithms are effective in detecting Alzheimer's disease early. limitation this takes a lot of time to locate the best features for given budget range.

In 2020,R, Saputra, et al [87]. The Particle Swarm Optimization (PSO) technique is used Use the Alzheimer OASIS 2 dataset of kaggle.com to test several decision tree algorithms with feature or characteristic selection. The result for studies utilizing 10-fold CV, via evaluating a decision tree approach to conducting the attribute and feature values, show that random forest(RF) method has the maximum degree of accuracy, with a value of 91.15 percent. The PSO method is used for feature selection, and the testing is frequent several times usage the (DT) algorithm, the Particle Swarm optimization based RF method has a kappa rate of 0.884 and precision value 93.56 percent. The challenges of limited sample numbers and low accuracy are the constraints of this paper. To boost accuracy, a combination of different feature selection approaches and local search methods is used.

In 2020,C, Park, et al [88]. The paper suggested the deep learning approach this uses (DNA) methylation data and large-scale gene expression (GE) to predict AD Modeling Alzheimer's disease using a multi-omics dataset is difficult since it requires integrating multiple omics data and dealing with large quantities of small-sample data. We came up with an innovative, yet simple, strategy to minimize the number of features in the multi-omics dataset based on differentially expressed genes and differentially methylated positions to address this issue. (AUC = 0.797, 0.756, 0.773, and 0.775, respectively). a list of the paper's limitations Highest computing speed possible.

In 2020 , K,Muhammed Niyas, et al [89]. suggest the efficient combination greedy searching and Fisher Score (FS) the selection for Alzheimer's diagnosis features. To classify Normal Controls, MCI the suggested technique achieves a 90% and 91% Balanced Classification Accuracy and then the Curve values 0.97/ 0.98 utilizing SVM, K-Nearest Neighbor, etc. The suggested technique provides greater sensitivity and specificity (84 percent and 82.5 percent, respectively). According to the results, the proposed

strategy for early Alzheimer's disease detection via effective feature selection is intriguing and may even be superior to present methods in some instances. Determining the criterion for the optimal combination of attributes based on ranking.

In 2021, N, Le, et al [90]. This work, our machine learning model was trained utilizing 35 expression characteristics using gene expression microarray data. The 35 – feature model outperformed classifiers by an average (AUC 98.3percent). The paper's limitations are due to the approach adopted, which is insufficient for predicting survival outcomes and even results in a prognosis that is polar opposite from the actual event.

Table 2 summarizes the most recent progress in the Alzheimer's disease prediction system (2011 - 2021), In focusing on the feature selection approach, it gives us a quick review of the work that has been done in this crucial medical domain.

Table 2. Summary of literature

References	Years	Dataset	Gene Selection Methods	Used Methods	Accuracy
B. BOOIJ, ET AL., [69]	2011	Not available to the general public.	Jack-knife (training data)	PLSR	ACC: 0.87 AUC: 0.94
L. Scheubert, et al.[70]	2012	The AD dataset GSE5281.	GA/SVM	NB	81.4
				C4.5	78.9
				KNN	87.0
				RF	87.0
				SVM+Gaussian kernel	85.7
				SVM + linear kernel	91.9
k, Lunnon, et al.[71]	2013	ANM1	RF, t-test with backward elimination and Meng score (training data)	Random Forest(RF)	ACC: 0.75
P,Johnson, et al.[72]	2014	Not public MCI/AD	GA	LR	AUC = 0.90
S, Sood, et al.[74]	2015	ANM1, ANM2 GEO:GSE60862)	(ULSAM Ageing), Bayesian statistic	KNN	AUC: 0.66 (ANM2) AUC: 0.73 (ANM1)
F, Sherif, et al [73]	2015	(WGS) data (ADNI)	PLINK	Supervised Bayesian network structural learning	66.13% and naive Bayes 59.43%
S, PAYLAKHI, ET AL.[75]	2015	GSE129	Fisher criteria	GA/SVM	100%
S,Zahra Paylakhi, et al.[76]	2016	(GSE1297)	Fisher criterion SAM GA	SVM	94.55
N, Voyle, et al.[77]	2016	ANM1 and ANM2	pick Size Tolerance (Training data) and REF	Random Forest(RF)	ACC: 70%
M, BARATI, ET AL .[78]	2016	GSE1297 GSE4757	Not described	NN+ MLp and perceptron	97%

		GSE28146 GSE32536 GSE6980			
M, BALAMURUGAN, ET AL [79].	2017	NACC having Data Set (RDD-UDS)	Not described	KNN	98.5
K, Nishiwaki, et al .[80]	2017	GDS810 GDS2795 GDS4135 GDS4136 GDS4758	RF	SVM-RFE	0.83
H, LI, ET AL .[81]	2017	GSE28490 GSE28491 GSE63060 GSE63061 GSE19151 GSE5281	Ref-REO	Not described	AUC: 0.77 (ANM1) And AUC: 0.73 (ANM2)
L, XU, ET AL[82]	2018	utilizing the protein sequence data supplied to forecast the onset of early-stage AD	Not described	SVM	85.7%
X, Li, et al.[83]	2018	The datasets were downloaded from GEO: GSE63060 and GSE63061.	LASSO	SVM	0.773
				RF	0.785
				RR	0.765
				Majority voting	
K, Sekaran, et al.[84]	2019	The dataset utilized is taken from the freely accessible information source specifically (GSE1297)	t-test	SVM	87.10
				NB	90.32
				RF	97.66
				MLP-NN	100
T, Lee, et al.[85]	2020	GEO:GSE63060), (GEO:GSE63061)	TF-related genes, VAE, the CFG scoring and hub genes	SVM,LR, RF,L1-GLM, and DNN	0.657, 0.874, and 0.804.
H ,Ahmed, et al [86]	2020	ADNI database	plink	(NB), (RF), (LR), and (SVM)	98.1%, 97.97%, 95.88%, and 83%, respectively
R, Saputra, et al.[87]	2020	The data on (https://www.oasis-brains.org/oasis)	PSO	ID3/C4.5/CHAID and RF	Auc (88.45%(c4.5) 89.54% (pso ID3) 89.54%(PSO CHAID) 91.15%(PSO Random Forest)
C, Park, et al.[88]	2020	GSE33000 and GSE44770 and GSE80970	naive Bayesian,	FNNs Exceeded SVM, and RF	Accuracy : 0.79, 0.75, 0.77, and 0.77.
K,Muhammed Niyas, et al.[89]	2020	ADNI-TADPOLE	FS and greedy searching algorithm.	SVM and KNN	84%, 82.5%

N, LE, ET AL. [90]	2021	GDS4602 and GDS460	RF	KNN, Naïve Bayes, RF, and SVM	95.5% 95.9 87% 98.3%

6 Discussion

As time goes on, more research on Alzheimer's Disease prediction utilizing various methodologies has been published. Comprehensive reviews of the current state of research and implementation are required because it is so important. As a result, the purpose of this study is to present a comprehensive overview of the most recent research in the field of Alzheimer's Disease detection that employs various techniques. From the end of 2011 until the present, there has been a lot of research on Alzheimer's Disease. Based on a review of the works under consideration, Because of the noisy data, feature extraction approaches were found to be far more suitable for automated identification of Alzheimer's Disease than feature selection techniques. Because the majority of biomedical datasets have noisy data rather than useless or redundant data. Feature selection is a tool that can be used to remove irrelevant and/or superfluous features in a variety of applications. There is no unique way of selecting features that can be used across all applications. Some techniques are used to remove unimportant characteristics while avoiding redundant features. A feature weighting algorithm based just on relevance does not adequately address the need for feature selection. Subset search algorithms look for candidate feature subsets based on an evaluation metric that measures how good each subset is.

The consistency measure and the association measure are two current evaluation tools that have been proved to be successful at eliminating both irrelevant and redundant characteristics. Experiments demonstrate that the number of iterations necessary to discover the optimum feature subset is usually at least quadratic to the number of features. As a result, existing subset search methods with quadratic or greater time complexity in terms of dimensionality do not have adequate scalability to deal with high dimensional data. Filters and wrappers are two types of feature selection strategies. Wrapper approaches typically outperform filter methods because the feature selection process is tailored to the classification technique being utilized. However, if there are a lot of features, they're usually too expensive to employ because every feature set should be evaluated with the trained model separately. Filter techniques are so much faster than wrapper methods, making them better suited to large data sets. To dealing with high-dimensional data, methods in a hybrid paradigm have recently been proposed to incorporate the benefits of both models. And there are just a few strategies for dealing with noisy data. As a preprocessing phase, feature extraction approaches have been proposed to reduce the impact of type noise on the learning process. According to research, the accuracy of classification achieved with various feature reduction algorithms is strongly dependent on the type of data. When opposed to approaches that discretely handle feature redundancy and/or irrelevant characteristics, techniques that handle both irrelevant and redundant features at the same time are far more robust and advantageous for the learning process. As a result, work based on a small amount of

data would not be labeled a significant addition to this discipline. The identification of Alzheimer's Disease using various approaches has three major limitations. The first is a data imbalance that can be addressed in future work by adding more features or knowledge-based characteristics to the model. The second issue was dealing with a large number of data; for this problem, cloud computing would be preferable to locally training a large amount of data, which would take more technical and manual effort. The last concern was a lack of available datasets, which is currently the most serious challenge in this field. There are a few reliable gene expression datasets for Alzheimer's Disease.

7 Conclusions

The detection of Alzheimer's Disease (AD), gene expression datasets and machine learning algorithms are commonly employed. Due to its vast dimensional features and small sample sizes, DNA microarray data present numerous hurdles to machine learning research. Features selection as a pre-processing method is only important in lowering the number of input features and in saving computing time and memory. Feature selection helps to improve classification accuracy. Researchers must also deal with the data's uneven distribution of classifications. A variety of test and training datasets have been located, but apart from the problem of using too many features for several small samples, the presence of outliers remains concern (i.e. dataset shift). Every year, researchers develop many new strategies to enhance earlier methods' classification accuracy and overcome limitations. Researchers also hope to assist biologists in discovering and understanding the fundamental pathway that connects gene expression to disease.

This challenge is being tackled through feature selection, and the results have been encouraging. Researchers are increasingly turning to hybrid feature selection strategies for guidance in their feature selection work. These approaches can essentially be categorized as a filter, wrapper, or embedding strategies. Given the enormous computer resources required by massive datasets, filtering algorithms are the most common. Wrapper and embedding techniques have been strategically avoided. These techniques have improved the robustness of the selected genes and the accuracy of the Alzheimer's Disease classification model.

8 References

[1] S. Mahajan, G. Bangar, and N. Kulkarni, "Machine Learning Algorithms for Classification of Various Stages of Alzheimer ' s Disease: A review," *Int. Res. J. Eng. Technol.*, vol. 07, no. 08, pp. 817–824, 2020.

[2] S. J. Mohammed, "A Proposed Alzheimer ' s Disease Diagnosing System Based on Clustering and Segmentation Techniques," *Eng. Technol.*, vol. 36, no. 2, pp. 160–165, 2018. https://doi.org/10.30684/etj.36.2B.12

[3] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick, "Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging," *Neurobiol.*

Aging, vol. 29, no. 4, pp. 514–523, 2008. https://doi.org/10.1016/j.neurobiolaging.2006.11.010

[4] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer{\textquoteright}s disease," *Alzheimer's {&} Dement. J. Alzheimer's Assoc.*, vol. 3, no. 3, pp. 186–191, 2007. https://doi.org/10.1016/j.jalz.2007.04.381

[5] T. Wang, R. G. Qiu, and M. Yu, "Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018. https://doi.org/10.1038/s41598-018-27337-w

[6] C. Davatzikos, S. M. Resnick, X. Wu, P. Parmpi, and C. M. Clark, "Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI," *Neuroimage*, vol. 41, no. 4, pp. 1220–1227, 2008. https://doi.org/10.1016/j.neuroimage.2008.03.050

[7] Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J; Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage. 2015 Jan 1;104:398-412. https://doi.org/10.1016/j.neuroimage.2014.10.002

[8] E. Pellegrini *et al.*, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 10, pp. 519–535, 2018. https://doi.org/10.1016/j.dadm.2018.07.004

[9] H. A. R. Akkar and S. A. Salman, "Detection of Biomedical Images by Using Bio-inspired Artificial Intelligent," *Eng. Technol. J.*, vol. 38, no. 2A, pp. 255–264. https://doi.org/10.30684/etj.v38i2A.319

[10] A. R. Abbas and A. O. Farooq, "Skin detection using improved ID3 algorithm," *Iraqi J. Sci.*, vol. 60, no. 2, pp. 402–410, 2019. http://dx.doi.org/10.24996/ijs.2019.60.2.20

[11] A. Abdulwahab, H. Attya, and Y. H. Ali, "Documents Classification Based On Deep Learning," *Int. J. Sci. Technol. Res.*, no. June, pp. 1–6, 2020.

[12] H. Bisgin *et al.*, "Comparing SVM and ANN based Machine Learning Methods for Species Identification of Food Contaminating Beetles," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018. https://doi.org/10.1038/s41598-018-24926-7

[13] W. M. S. Abedi, I. Nadher, and A. T. Sadiq, "Modified deep learning method for body postures recognition," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 2, pp. 3830–3841, 2020.

[14] A. Karim, A. Hassan, and M. Alawi, "Proposed Handwriting Arabic Words classification Based On Discrete Wavelet Transform and Support Vector Machine," *Iraqi J. Sci.*, vol. 58, no. 2C, 2017. https://doi.org/10.24996/ijs.2017.58.2C.19

[15] N. J. N. N. P. S. Daniel E Shumer, "乳鼠心肌提取 HHS Public Access," *Physiol. Behav.*, vol. 176, no. 12, pp. 139–148, 2017.

[16] A. T. Sadiq and N. H. Shukr, "Classification of Cardiac Arrhythmia using ID3 Classifier Based on Wavelet Transform," *Iraqi J. Sci.*, vol. 54, no. 4, pp. 1167–1175, 2013.

[17] Y. Zhang, J. M. Gorriz, and Z. Dong, "Deep learning in medical image analysis," *J. Imaging*, vol. 7, no. 4, p. NA, 2021. https://doi.org/10.3390/jimaging7040074

[18] S. G. Yilmaz and S. Arslan, "Effects of progressive relaxation exercises on anxiety and comfort of Turkish breast cancer patients receiving chemotherapy," *Asian Pacific J. Cancer Prev.*, vol. 16, no. 1, pp. 217–220, 2015. https://doi.org/10.7314/APJCP.2015.16.1.217

[19] N. F. Hassan and H. I. Abdulrazzaq, "Pose invariant palm vein identification system using convolutional neural network," *Baghdad Sci. J.*, vol. 15, no. 4, pp. 502–509, 2018. https://doi.org/10.21123/bsj.2018.15.4.0502

[20] A. T. Sadiq and S. M. Abdullah, "Hybrid intelligent technique for text categorization," *Proc. - 2012 Int. Conf. Adv. Comput. Sci. Appl. Technol. ACSAT 2012*, vol. 2, no. 2, pp. 238–245, 2012. https://doi.org/10.1109/ACSAT.2012.50

[21] C. Plant *et al.*, "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease," *Neuroimage*, vol. 50, no. 1, pp. 162–174, 2010. https://doi.org/10.1016/j.neuroimage.2009.11.046

[22] Liu M, Zhang D, Shen D; Alzheimer's Disease Neuroimaging Initiative. Ensemble sparse classification of Alzheimer's disease. Neuroimage. 2012 Apr 2;60(2):1106-16. https://doi.org/10.1016/j.neuroimage.2012.01.055

[23] D. G. Clark *et al.*, "Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 2, pp. 113–122, 2016. https://doi.org/10.1016/j.dadm.2016.02.001

[24] H. Motieghader, A. Najafi, B. Sadeghi, and A. Masoudi-Nejad, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata," *Informatics Med. Unlocked*, vol. 9, no. October, pp. 246–254, 2017. https://doi.org/10.1016/j.imu.2017.10.004

[25] H. Motieghader, S. Gharaghani, Y. Masoudi-Sobhanzadeh, and A. Masoudi-Nejad, "Sequential and mixed genetic algorithm and learning automata (SGALA, MGALA) for feature selection in QSAR," *Iran. J. Pharm. Res.*, vol. 16, no. 2, pp. 533–553, 2017. https://dx.doi.org/10.22037/ijpr.2017.2108

[26] Q. Li, X. Wu, L. Xu, K. Chen, L. Yao, and R. Li, "Multi-modal discriminative dictionary learning for Alzheimer's disease and mild cognitive impairment," *Comput. Methods Programs Biomed.*, vol. 150, pp. 1–8, 2017. https://doi.org/10.1016/j.cmpb.2017.07.003

[27] C. Park, Y. Yoon, O. Min, S. J. Yu, and J. Ahn, "Systematic identification of differential gene network to elucidate Alzheimer's disease," *Expert Syst. Appl.*, vol. 85, pp. 249–260, 2017. https:/doi.org/10.1016/j.eswa.2017.05.042

[28] G. Meng, X. Zhong, and H. Mei, "A systematic investigation into Aging Related Genes in Brain and Their Relationship with Alzheimer's Disease," *PLoS One*, vol. 11, no. 3, pp. 1–17, 2016. https://doi.org/10.1371/journal.pone.0150624

[29] J. Liu, X. Wang, Y. Cheng, and L. Zhang, "Tumor gene expression data classification via sample expansionbased deep learning," *Oncotarget*, vol. 8, no. 65, pp. 109646–109660, 2017. https://doi.org/10.18632/oncotarget.22762

[30] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Distributed feature selection: An application to microarray data classification," *Appl. Soft Comput. J.*, vol. 30, pp. 136–150, 2015. https://doi.org/10.1016/j.asoc.2015.01.035

[31] N. Jameel and H. S. Abdullah, "A Proposed Intelligent Features Selection Method Using Meerkat Clan Algorithm," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, 2021. https://doi.org/10.1088/1742-6596/1804/1/012061

[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[33] A. T. Sadiq and K. S. Musawi, "Modify Random Forest Algorithm Using Hybrid Feature Selection Method," *Int. J. Perceptive Cogn. Comput.*, vol. 4, no. 2, pp. 1–6, 2018. https://doi.org/10.31436/ijpcc.v4i2.59

[34] A. T. Sadiq and S. A. Chawishly, "Intelligent Methods to Solve Null Values Problem in Databases Intelligent Methods to Solve Null Values Problem in Databases," *J. Adv. Comput. Sci. Technol. Res.*, no. December, 2016.

[35] L. Li *et al.*, "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, 2005. https://doi.org/10.1016/j.ygeno.2004.09.007

[36] P. Di Lena, K. Nagata, and P. Baldi, "Deep spatio-temporal architectures and learning for protein structure prediction," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 512–520, 2012.

[37] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Inf. Fusion*, vol. 50, pp. 158–167, 2019. https://doi.org/10.1016/j.inffus.2018.11.019

[38] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013. https://doi.org/10.1007/s10115-012-0487-8

[39] S. Liang, A. Ma, S. Yang, Y. Wang, and Q. Ma, "A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 88–97, 2018. https://doi.org/10.1016/j.csbj.2018.02.005

[40] E. Hancer, B. Xue, D. Karaboga, and M. Zhang, "A binary ABC algorithm based on advanced similarity scheme for feature selection," *Appl. Soft Comput. J.*, vol. 36, pp. 334–348, 2015. https://doi.org/10.1016/j.asoc.2015.07.023

[41] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, pp. 1200–1205, 2015. https://doi.org/10.1109/MIPRO.2015.7160458

[42] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci. (Ny).*, vol. 282, pp. 111–135, 2014. https://doi.org/10.1016/j.ins.2014.05.042

[43] F. Rafii, B. D. R. Hassani, and M. A. Kbir, "New approach for microarray data decision making with respect to multiple sources," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1294, 2017. https://doi.org/10.1145/3090354.3090463

[44] L. H. Augenlicht, J. Taylor, L. Anderson, and M. Lipkin, "Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 8, pp. 3286–3289, 1991. https://doi.org/10.1073/pnas.88.8.3286

[45] C. S. Kong, J. Yu, F. C. Minion, and K. Rajan, "Identification of biologically significant genes from combinatorial microarray data," *ACS Comb. Sci.*, vol. 13, no. 5, pp. 562–571, 2011. https://doi.org/10.1021/co200111u

[46] Q. K. Kadhim, "Classification of Human Skin Diseases using Data Mining," *Int. J. Adv. Eng. Res. Sci.*, vol. 4, no. 1, pp. 159–163, 2017. https://doi.org/10.22161/ijaers.4.1.25

[47] D. M. Mutch, A. Berger, R. Mansourian, A. Rytz, and M. A. Roberts, "Microarray data analysis: a practical approach for selecting differentially expressed genes.," *Genome Biol.*, vol. 2, no. 12, 2001. https://doi.org/10.1186/gb-2001-2-12-preprint0009

[48] S. Selvaraj and J. Natarajan, "Microarray Data Analysis and Mining Tools," vol. 6, no. 3, 2011. https://doi.org/10.6026/97320630006095

[49] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017. https://doi.org/10.1016/j.ygeno.2017.01.004

[50] O. D. Madeeh and H. S. Abdullah, "An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, 2021. https://doi.org/10.1088/1742-6596/1804/1/012008

[51] Y. Wang *et al.*, "Gene selection from microarray data for cancer classification - A machine learning approach," *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005. https://doi.org/10.1016/j.compbiolchem.2004.11.001

[52] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[53] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014. https://doi.org/10.1007/s00521-013-1368-0

[54] L. Ladha and T. Deepa, "Feature Selection Methods And Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 1787–1797, 2011, [Online]. Available: http://journals.indexcopernicus.com/abstract.php?icid=945099

[55] N. Jameel and H. S. Abdullah, "Intelligent Feature Selection Methods : A Survey," *Eng. Technol. J.*, vol. 39, no. 01, pp. 175–183, 2021. https://doi.org/10.30684/etj.v39i1B.1623

[56] M. Veerabhadrappa and L. Rangarajan, "Bi-level dimensionality reduction methods using feature selection and feature extraction," *Int. J. Comput. Appl.*, vol. 4, no. 2, pp. 33–38, 2010. https://doi.org/10.5120/800-1137

[57] A. Tahseen Ali, H. S. Abdullah, and M. N. Fadhil, "Voice recognition system using machine learning techniques," *Mater. Today Proc.*, no. xxxx, 2021. https://doi.org/10.1016/j.matpr.2021.04.075

[58] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings, Twent. Int. Conf. Mach. Learn.*, vol. 2, pp. 856–863, 2003.

[59] S. D. Khudhur and A. K. Taqi, "Edge Detection and Features Extraction for Dental X-Ray," *Eng. &Tech.Journal*, vol. 34, no. September, pp. 2420–2432, 2016.

[60] S. S. Al-rawi, A. T. Sadiq, and W. M. Alaluosi, "Feature Extraction of Human Facail Expressions Using Haar Wavelet and Neural network," *Iraqi J. Sci.*, vol. 57, no. 2, pp. 1558–1565, 2016.

[61] A. Masood *et al.*, "Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images," *J. Biomed. Inform.*, vol. 79, no. June 2017, pp. 117–128, 2018. https://doi.org/10.1016/j.jbi.2018.01.005

[62] H. N. Abdullah, "Deep CNN Based Skin Lesion Image Denoising and Segmentation using Active Contour Method," *Eng. Technol. J.*, vol. 37, no. 11, 2019. https://doi.org/10.30684/etj.37.11A.3

[63] A. A. Abdulhussein and F. A. Raheem, "Hand Gesture Recognition of Static Letters American Sign Language (ASL) Using Deep Learning," *Eng. Technol. J.*, vol. 38, no. 6A, pp. 926–937, 2020. https://doi.org/10.30684/etj.v38i6A.533

[64] I. S. Abed, "Lung Cancer Detection from X-ray images by combined Backpropagation Neural Network and PCA," *Eng. Technol. J.*, vol. 37, no. 05, 2019. https://doi.org/10.30684/etj.37.5A.3

[65] S. Taha Ahmed and S. Malallah Kadhem, "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 16, p. 95, 2021. https://doi.org/10.3991/ijim.v15i16.24191

[66] T. Howley, M. G. Madden, M. O. Connell, and A. G. Ryder, "Applications and Innovations in Intelligent Systems XIII," *Appl. Innov. Intell. Syst. XIII*, no. August 2014, 2006. https://doi.org/10.1007/1-84628-224-1

[67] A. E. Guyon, Isabelle, "An Introduction to Variable and Feature Selection," *Anal. Chim. Acta*, vol. 703, no. 2, pp. 152–162, 2011. https://doi.org/10.1016/j.aca.2011.07.027

[68] H. A. R. Akkar and S. Q. Hadad, "Diagnosis of Lung Cancer Disease Based on Back-Propagation Artificial Neural Network Algorithm," *Eng. Technol. J.*, vol. 38, no. 3B, pp. 184–196, 2020. https://doi.org/10.30684/etj.v38i3B.1666

[69] B. B. Booij *et al.*, "A gene expression pattern in blood for the early detection of Alzheimer's disease," *J. Alzheimer's Dis.*, vol. 23, no. 1, pp. 109–119, 2011. https://doi.org/10.3233/JAD-2010-101518

[70] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen, "Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC Bioinformatics*, vol. 13, no. 1, 2012. https://doi.org/10.1186/1471-2105-13-266

[71] K. Lunnon *et al.*, "A blood gene expression marker of early Alzheimer's disease," *J. Alzheimer's Dis.*, vol. 33, no. 3, pp. 737–753, 2013. https://doi.org/10.3233/JAD-2012-121363
[72] P. Johnson *et al.*, "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease," *BMC Bioinformatics*, vol. 15, no. Suppl 16, pp. 1–14, 2014. https://doi.org/10.1186/1471-2105-15-S16-S11
[73] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks," *Adv. Bioinformatics*, vol. 2015, pp. 1–9, 2015. https://doi.org/10.1155/2015/639367
[74] S. Sood *et al.*, "A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status," *Genome Biol.*, vol. 16, no. 1, pp. 1–17, 2015. https://doi.org/10.1186/s13059-015-0750-x
[75] S. Z. Paylakhi, S. Ozgoli, and S. H. Paylakhi, "A novel gene selection method using GA/SVM and Fisher criteria in Alzheimer's disease," *ICEE 2015 - Proc. 23rd Iran. Conf. Electr. Eng.*, vol. 10, pp. 956–959, 2015. https://doi.org/10.1109/IranianCEE.2015.7146349
[76] S. Zahra Paylakhi, S. Ozgoli, and S. Paylakhi, "Identification of Alzheimer disease-relevant genes using a novel hybrid method," *Prog. Biol. Sci.*, vol. 6, no. 1, pp. 37–46, 2016, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed
[77] N. Voyle *et al.*, "A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis," *J. Alzheimer's Dis.*, vol. 49, no. 3, pp. 659–669, 2015. https://doi.org/10.3233/JAD-150440
[78] M. Barati and M. Ebrahimi, "Identification of Genes Involved in the Early Stages of Alzheimer Disease Using a Neural Network Algorithm," *Gene, Cell Tissue*, vol. 3, no. 3, 2016. https://doi.org/10.17795/gct-38415
[79] M. Balamurugan, A. Nancy, and S. Vijaykumar, "Alzheimer's disease diagnosis by using dimensionality reduction based on KNN Classifier," *Biomed. Pharmacol. J.*, vol. 10, no. 4, pp. 1823–1830, 2017. https://doi.org/10.13005/bpj/1299
[80] K. Nishiwaki, K. Kanamori, and H. Ohwada, "Gene Selection from Microarray Data for Alzheimer's Disease Using Random Forest," *Int. J. Softw. Sci. Comput. Intell.*, vol. 9, no. 2, pp. 14–30, 2017. https://doi.org/10.4018/IJSSCI.2017040102
[81] H. Li *et al.*, "Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017. https://doi.org/10.1038/s41598-017-13700-w
[82] Ei, G. Liang, C. Liao, G. Den Chen, and C. C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, 2018. https://doi.org/10.3390/molecules23123140
[83] Inzhong *et al.*, "Systematic Analysis and Biomarker Study for Alzheimer's Disease," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, 2018. https://doi.org/10.1038/s41598-018-35789-3
[84] K. Sekaran and M. Sudha, "Diagnostic gene biomarker selection for alzheimer's classification using machine learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 2348–2352, 2019. https://doi.org/10.35940/ijitee.L3372.1081219
[85] T. Lee and H. Lee, "Prediction of Alzheimer's disease using blood gene expression data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020. https://doi.org/10.1038/s41598-020-60595-1
[86] H. Ahmed, H. Soliman, and M. Elmogy, "Early Detection of Alzheimer's Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, pp. 15–20, 2020. https://doi.org/10.1109/ICDABI51230.2020.9325640

[87] R. A. Saputra *et al.*, "Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020. https://doi.org/10.1088/1742-6596/1641/1/012025

[88] 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. https://doi.org/10.1016/j.eswa.2019.112873

[89] K. P. Muhammed Niyas and P. Thiyagarajan, "Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021. https://doi.org/10.1016/j.jksuci.2020.12.009

[90] N. Q. K. Le, D. T. Do, T.-T.-D. Nguyen, N. T. K. Nguyen, T. N. K. Hung, and N. T. T. Trang, "Identification of gene expression signatures for psoriasis classification using machine learning techniques," *Med. Omi.*, vol. 1, no. December 2020, p. 100001, 2021. https://doi.org/10.1016/j.meomic.2020.100001

9 Authors

Shaymaa Taha Ahmed M.Sc. (2015) in (India), currently a postgraduate student studying PhD in Department of Computer Sciences, University of Technology, Baghdad, Iraq. Affiliation: University of Diayla Dept.: computer science / College: basic of education Specialization: - Computer science \ information system. Research Interests: Cloud Computing-Deep Learning-Machine learning-AI -Data mining Google Site: Google scholar (email: cs.19.25@grad.uotechnology.edu.iq, Shaymaa.taha.ahmed@basicedu.uodiyala.edu.iq & mrs.sh.ta.ah@gmail.com, ORCID: https://orcid.org/0000-0002-4986-2475).

Suhad Malallah Kadhem PhD in Computer Sciences/ 2003/ Computer Science Department/ University of Technology. Scientific Specialization: Natural Language Processing. Scientific Title: Assistant Professor/ 20012. Scientific Research Interest: Natural Language Processing, Machine translation, Artificial intelligence, Information Security, Information Hiding, machine learning, deep learning and data mining (email: 110102@uotechnology.edu.iq).

Article submitted 2021-09-26. Resubmitted 2022-01-12. Final acceptance 2022-01-15. Final version published as submitted by the authors.