DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

Predicting Alzheimer's Disease Using Filter Feature Selection Method

Shaymaa Taha Ahmed[1], Suhad Malallah Kadhem[2]
[1]Department of Computer Science, College of Basic Education, University of Diyala, Diyala, Iraq
[2]Computer Sciences Department, University of Technology, Baghdad, Iraq
[1]110102@uotechnology.edu.iq, [2]mrs.sh.ta.ah@gmail.com

Abstract—*Alzheimer's disease (AD) is caused by multiple variables. Alzheimer's disease development and progression are influenced by genetic variants. The molecular pathways causing Alzheimer's disease are still poorly understood. In Alzheimer's disease research, determining an effective and reliable diagnosis remains a major difficulty, particularly in the early stages (i.e., Moderate Cognitive Impairment (MCI)). Researchers and technologists working in the fields of machine learning and data mining can help improve the situation, early AD diagnosis but face a hurdle when it comes to high-dimensional data processing. By reducing irrelevant and redundant data from microarray gene expression data, the technique of feature selection can save computing time, improve learning accuracy, and encourage a deeper effect on the learning system or data. The feature selection strategy described in this article reduces data noise well. In particular, Pearson's correlation coefficient is used to assess data redundancy. The efficacy of these features is assessed using the Support Vector Machine (SVM) classification approach. The proposed approach has an accuracy of up to 91.1 %. As a result, newly established approaches for early diagnosis of Alzheimer's disease(AD) are being improved.*

Index Terms— *Alzheimer's Disease, Support vector machine, machine learning, feature selection, Pearson's correlation coefficient.*

I. INTRODUCTION

The most frequent type of dementia in the elderly is Alzheimer's disease, characterized by a decrease in the number of neurons and their connections throughout time [1]. According to reports, In the next 20 years, the number of people with Alzheimer's disease will triple, with 1 in 85 individuals affected by 2050[2]. Because Alzheimer's disease is slow-progressing dementia with first pathology transformation decades before memory loss emerges, it is difficult to predict when it will strike, a precise and timely diagnosis of the disease, particularly in its early stages, is critical for timely therapies. The molecular architecture of Alzheimer's disease, like that of most diseases, is displayed at various levels, transcriptional in genomics, posttranscriptional, as well as epigenetic changes [3]. Currently, Alzheimer's disease is diagnosed through clinical screening, such as imaging techniques or cerebrospinal fluid examination [4]. Early-stage dementia typically leads to erroneous diagnosis and, as a result, delays in receiving appropriate therapies. As a result, the development of useful and efficient biomarkers capable of establishing accurate correspondences and correlations with clinical symptoms has become a top priority [5][6].

Gene selection is a technique for determining the best set of genes (features) that influence label classification the most [7][8]. When working with large datasets, it's

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

important to remember that the more dimensions the better, where a dimension exceeds the number of samples, gene selection is very crucial [9][10]. In general, a large number of dimensions can reduce categorization accuracy. Even though it does not affect accuracy, it may hurt performance [11][12].

Machine learning is rapidly evolving, and the amount of data collected via the internet is increasing, traditional data analysis approaches are becoming increasingly challenging to adapt to today's big data concerns, prompting the creation of numerous data preparation techniques[13][14]. It is the process of picking a subset of features (variables, characteristics) from a huge dataset with thousands or tens of thousands of features that satisfy the analytic goal. Analysts may make use of many benefits, improved predictive model performance, as well as increased efficiency data processing, are just a few of the benefits[15][16].

Feature selection has the following advantages:

(a) decreases the dataset's dimension, lowering the cost of computer resources.

(b) Reduces data noise, which enhances classification model performance.

(c) making it easier to visualize and comprehend data.

The basic goal of generic feature selection is to identify a group of related qualities that are associated with certain events or phenomena.

In this research, New redundancy and relevance metrics are available. The Pearson correlation coefficient (PCC) is used as a measure of redundancy and relevance[17]. It's a metric for analyzing global feature correlations, according to the description. The following is a summary of this paper's contributions:

- Seeing if our definition of PCC as an optimization algorithm is correct, PCC and other feature subset evaluation methodologies were subjected to a correlation coefficient investigation.
- The PCC is a novel metric that measures the entire relevance and redundancy of a feature subset. After that, the characteristics of PCC are discussed.
- The PCC of feature subsets is calculated using the correlation among two features as the calculating.
- We examine which sets of data are covered by the method.

The rest of this work is organized in the following manner: Section II discusses the relevant work, which is mostly concerned with comparative approaches. The approaches for selecting features are described in Section III. Theoretical foundations of our materials and methods are described in sections IV and V. The research methodology and discussion and result are summarized in Sections VI,VII and VIII and other future research aims are outlined.

II. RELATED WORK

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

In 2011, B. Booij *et al.* [18], 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 55/63. (AUC 87 %).

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

In 2012, L. Schubert, *et al.* [19]. 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 of the lesser sample sizes in addition to the unstable nature of this algorithm being presented.

In 2013,k, Lunnon, *et al*. [20] ,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*. [21]. 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 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.* [3], 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 the advantage of an 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 of 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*. [22]. To predict HC conversion to MCI/AD, we used Bayesian statistics (ULSAM Ageing) and KNN with an 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*. [23]. 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 genes eliminated. A (GA-SVM) then used to choose distinct subsets of maximally informative genes using different training sets. The Fisher Score and (GA)(SVM) approaches that combined for a 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 be erroneous or else missing.

In 2016 , S,Zahra Paylakhi, *et al*. [24]. These methods combine the fisher Score, significant analysis of microarrays, and a (GA)- (SVM). A Fisher technique is employed to removing 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 used. Microarray data from AD patients were used to test the proposed technique. The result appears that the 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 to 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*. [25]. 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.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

In 2016,M, Barati, *et al*. [26]. 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*. [27]. The 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 the clinical and statistical datasets. The drawbacks of the KNN method are based on the feature from 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*. [7]. 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 of all datasets used are microarrays, hence their RNA-seq application is more accurate and less noisy.

In 2017,H, Li, *et al*. [28]. 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*. [29]. 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 the predictions method.

In 2018,X, Li, *et al.* [30]. In this paper, the 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*. [31]. 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 the 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 used 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)

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

and normal genes, demonstrating its usefulness. The study's weakness is that the training set is possible to contain a large amount of noise, which could have an impact on model performance.

In 2020,T, Lee, *et al*. [32]. For the aim of this research. Five (5) feature selection approaches and five classifications have been used to identify genes related to 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 , K,Muhammed Niyas, *et al*. [33]. suggest the efficient combination of 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 2020, H,Ahmed, *et al*. [4] . The focus of this research is on the use of ML approaches to identify AD biomarkers. Random Forest (RF), Naive Bayes (NB), (LR) , and Support Vector Machine algorithms were used to every Alzheimer’s disease genetic information from ADNI-1 imaging project datasets. Naive 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 a given budget range.

In 2020,R, Saputra, *et al*. [34]. The Particle Swarm Optimization (PSO) technique is used Use the Alzheimer OASIS 2 dataset of kaggle.com to test several decision tree algorithms with the feature or characteristic selection. The result for studies utilizing 10-fold CV, via evaluating a decision tree approach to conducting ([70] Sekaran)the attribute and feature values, show that the 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 using the (DT) algorithm, the Particle Swarm optimization based RF method has a kappa rate of 0.884 and a precision value of 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*. [35]. 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

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

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 2021, N, Le, *et al*. [36]. IN this work, our machine learning model was trained to utilize 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.

III. TECHNIQUES FOR FEATURE SELECTION

In this section, we looked at different ways of selecting features . Depending on the method used to locate the necessary features, Filtering strategies and wrapping approaches are typically used to separate this feature selection[37][38]. Filtering approaches evaluate the significance of characteristics by analyzing just the data's intrinsic attributes [39][40].

Most of the time, the relevance scores between each feature and the class vector are calculated, as well as the highest-scoring features are selected. Filtering techniques are basic, quick, and simple to comprehend[41][42]. They do not, however, take into account redundancy or the interaction of characteristics; instead, they believe these features are unrelated. To capture feature interactions, wrapper approaches incorporate a classification model into the evaluation of feature subsets. Because the number of features increases exponentially, the space of feature subsets expands exponentially, to direct the search toward an optimal subset, heuristic search methods such as forward search and backward elimination are used [43][44]. There are three different ways to select features: unsupervised, supervised, and semi-supervised [45][46]. When evaluating the significance of features without labels, unsupervised feature selection algorithms may use data variance or data distribution, supervised feature selection methods, on the other hand, examine the importance of features by evaluating their correlation with the classification method[47][48]. To improve unsupervised feature selection, methods for selecting features that are semi-supervised add more information, employ a small amount of labeled data [20][49]. Methods for selecting features can be based on statistics[50], information theory [51], manifold [52], and rough set[53], and can be classified according to various criteria, according to the theoretical concept.

There are three types of feature selection algorithms: supervised, unsupervised, or semi-supervised based on the type of data utilized for training (labeled, unlabeled, or partially labeled). A unified architecture for supervised, unsupervised, and semi-supervised feature selection is shown in *Fig. 1*.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

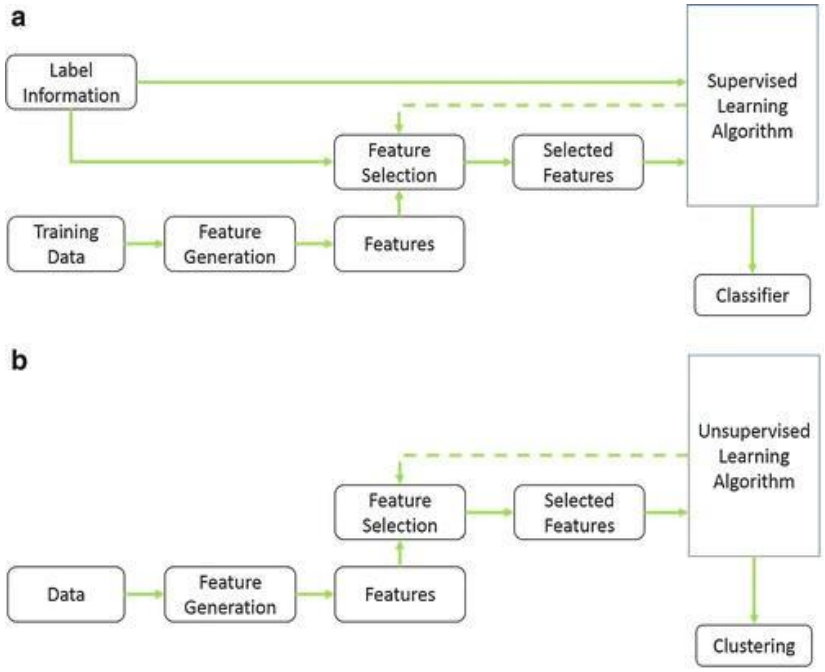

FIG. 1.A FRAMEWORK FOR SELECTION FEATURES.

Filter, wrapper, and embedding models are the three types of feature selection approaches based on their relationship to learning methods. Feature selection techniques can be constructed using correlation, Euclidean distance, consistency, dependence, and information measures, depending on the assessment criterion. Feature selection methods are classified as forwarding increase, backward deletion, random, or hybrid models, depending on the search process. There are two types of output from feature selection methods: feature rank (weighting) and subset selection models.

The filter model only considers the link between a feature and the class label. The wrapper model has a lower computational cost. The filter model's assessment criterion is crucial. Meanwhile, during the learning model's training process, the embedded model [54] selects a feature, the feature selection result is automatically outputted once the training procedure is completed [55], as shown in *Fig. 2*.

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.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

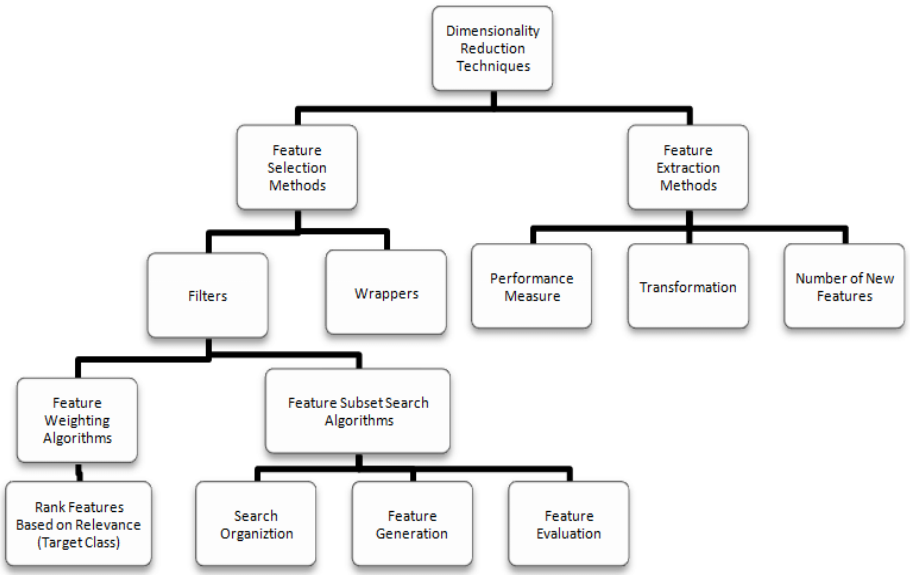

FIG. 2. DIMENSIONALITY REDUCTION APPROACHES: A HIERARCHICAL STRUCTURE.

IV. DATA COLLECTION

The NCBI-GEO database provided the microarray gene expression high-throughput datasets in AD [20]. The results of two different studies mononuclear cells from the peripheral blood taken from healthy people aged controls GSE132903 and GSE33000, respectively, were deposited in the NCBI-GEO database and analyzed utilizing human MGC cDNA microarrays in AD and non-AD patients, respectively[56]. GSE132903 comprised samples from a total of 195 people, including 97 controls and 98 AD patients[57] , who looked for novel biomarkers using DNA and blood protein investigations. 624 blood samples were examined by GSE33000 (314 from controls, 310 from AD patients). Who used the NCBI-GEO2R web tool to find DEGs in AD patients compared to age-matched controls. In our work, the gene expression dataset was first normalized using the log2 transformation . Datasets were then analyzed with NCBI's GEO2R program, with Limma for hypothesis testing and the Benjamini & Hochberg correction for false discovery rate control. A cut-off threshold for detecting significant DEGs was a p-value of less than.01. To detect shared DEGs from the two datasets, a Venn analysis was done using the free application Jvenn [58].

V. PROPOSED METHODS

The proposed framework consists of data pre-processing, feature reduction and classification.

A. Data Pre-processing

We use GSE13290 and GSE33000[10] blood dataset and map the probe set IDs to gene official symbols according to the GPL570 annotation table. The dataset contains 161 samples out of which 87 are diagnosed with Alzheimer's disease and 74 samples are from the healthy control group. Totally, there are 24,438 unique gene symbols. We split the analysis of the dataset into two parts. They are:

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

• Analyzing the performance of classification algorithms

B. Feature Selection

Pearson's Correlation Coefficient (PCC), a heuristic filter approach, is suggested in this study. The presented methods aim to enhance accuracy outcomes across binary medical datasets by selecting the smallest feature subset possible[59]. Furthermore, because both proposed methods use a filter approach, the feature subset would be chosen based on a satisfactory run time. Both strategies rely on natural data specifications inside two-class datasets. The framework for early Alzheimer's disease prediction as shown in *Fig. 3*.

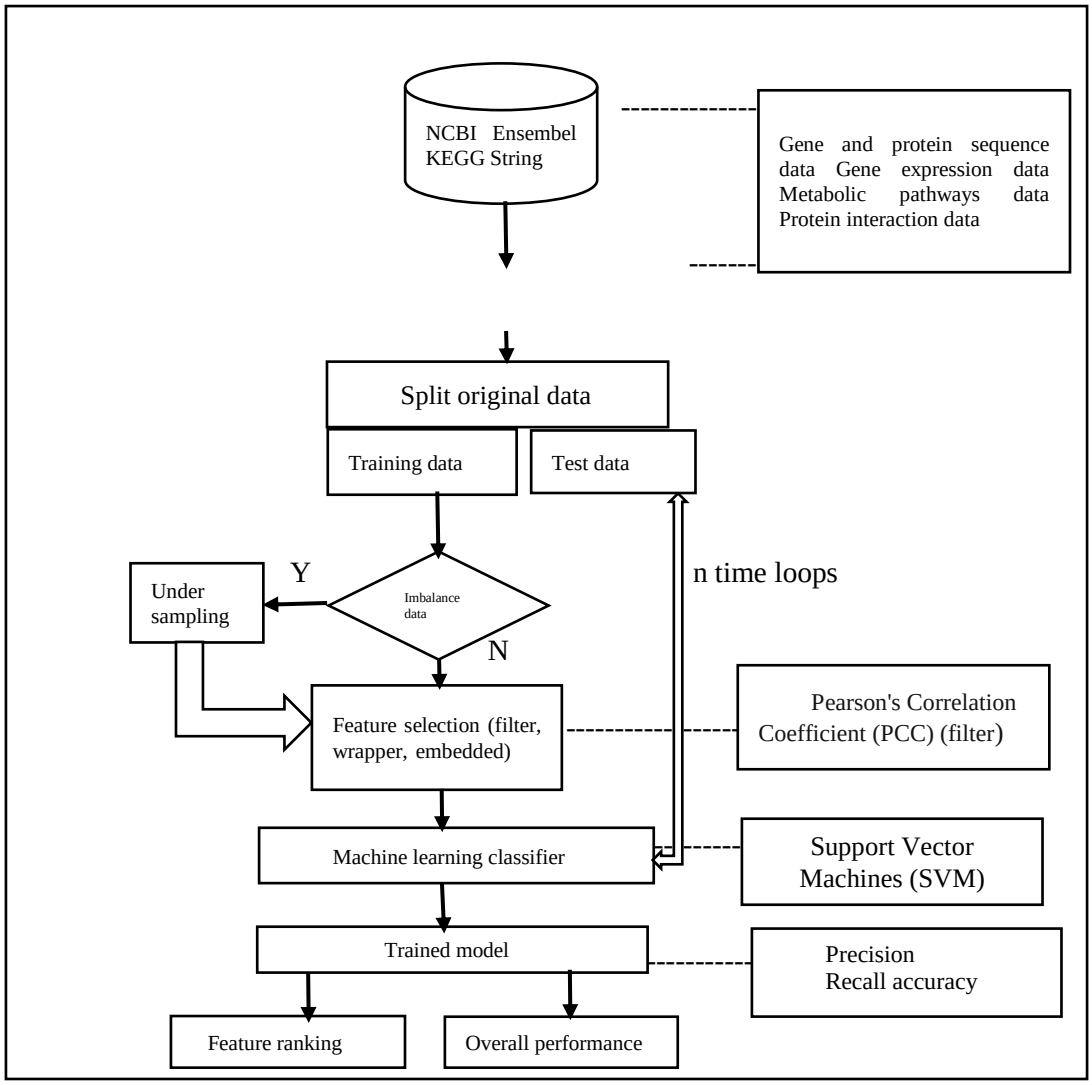

FIG. 3. A SIMPLE ILLUSTRATION OF MACHINE LEARNING AND CLASSIFICATION STEPS CAN BE USED TO DIAGNOSE ALZHEIMER'S DISEASE.

-Pearson's Correlation Coefficient (PCC)

PCC calculates the correlation between the weights of the specified subgroups of the attributes [60]. The result of PCC is in the range [1, 1]. 1 denotes the complete correlation between weight vectors, while 1 denotes anti-correlation between weight vectors. There is no correlation between the weight vectors when the value is 0. When the weight approaches zero to have a wider number of characteristics, it suggests that the system is more stable. The symmetric stability measure is known as PCC.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

$$PCC(W_i, W_j) = \frac{\sum(wit-\mu Wi)(Wjt-\mu Wj)}{\sqrt{\sum(Wit-\mu Wi)(Wjt-\mu Wj)}} \quad (1)$$

$where\ W =$ Mean of the feature set f

To get ranks, various feature selection techniques were applied to a dataset[61], subsets of features, and weights. Because it uses actual feature coefficients, stability based on feature rankings (SR). Sw provides the highest level of steadiness. The value of stability using subsets (SS) does not correlate with the two other measures. The higher the number of occurrences of the feature subsets chosen, the better, the more likely there are to be featured in common. As a result, the stability value will rise.

Offer a unique stability estimator that meets all of the desired features of a stability measure[62]. The following is the formula for this innovative stability metric:

$$\Phi(z)=1-\frac{\frac{1}{m}\sum m\ fi}{klm(1-klm)}$$

$$f^2=\frac{N}{N-1}P_f(1-P_f) \quad (2)$$

k=Average number of feature select over the N feature set in z.

f^2=sample variance of z_f.

C. Classification Method

To now, data mining technologies have been utilized to assess the early identification of Alzheimer's disease. In this study, approaches based on SVM were found to be one of the most extensively utilized ways for detecting early Alzheimer's disease[63].A Support Vector Machines (SVM): Its major goal is to make the data easier to understand for the user. SVM is used to appropriately separate the data of the two classes[64].SVM proved to be the most accurate of the ML models tested.

SVM is a directed study model that classifies by separating the objects using a hyperplane. It can be used for both classification and regression. The hyperplanes are drawn with the help of the margins. The main goal is to maximize the distance between the hyperplane and the margin.

The margins are drawn with the help of support vectors that are belonging to the objects. The main advantage of SVM is that it can distinguish linear and non-linear objects. *Fig. 1* shows the steps in predicting Alzheimer disease using machine learning algorithms.

classifier = svm (formula=age, visit, MMSE, EDUC .,data = train, type = 'C-classification', kernel = 'linear')

The required packages for the SVM classifier in R are caret and e1071 packages. The formula consists of the fields that are considered for prediction. The basic type c-classification and the linear kernel are chosen. They both mostly depend on the data used[65].

The psychological parameters are given as the input for the classifier. When the classifier is trained and given for testing, it predicts the output with an accuracy of 91.1%.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

VI. RESULTS

We used univariate selection which is a statistical test that could be used to select those features that have the strongest relationship with the output variable (AD or non AD), to reduce thousands of differential genes upto small subsets as it works better with larger datasets. We selected 200 features with the highest scores for further analysis.

VII. DISCUSSION

The diagnosis of Alzheimer's disease(AD) is now based on neuropsychological testing and neuroimaging, but finding reliable and precise biomarkers for diagnosis and prognosis remains a difficulty . We explored gene expression patterns seen in peripheral blood using a systems biology approach in this study cells of Alzheimer's patients and discovered many of promising candidates molecular targets that could be used as biomarkers for the disease. As a result, our findings may shed light on the path to Alzheimer's disease.

Microarray data is widely used in biomedical research and has emerged as a valuable tool for discovering potential biomarker candidates . DEGs are frequently identified using microarray gene expression profiling in a variety of illnesses, including Alzheimer's disease . In two transcriptomic datasets, analysis of gene expression patterns in the blood of Alzheimer's patients revealed significant changes in gene expression profiles. Over-representation analysis in the ribosome and complement systems showed AD and neurodegeneration-related molecular pathways. The many kinds of feature selection approaches and feature selection strategies were summarized. Because they both employ their learning process for feature selection, features selected by wrapper-based and embedding techniques may do not work well with other classifiers. Filter-based techniques have less computational complexity than embedded and wrapper-based techniques. Wrapper-based approaches have a considerable risk of overfitting due to their complexity. Due to their resistance to overfitting, filter-based approaches produce more stable sets of selected features. Because the curse of dimensionality makes stable feature selection difficult, advancement in the related discipline will lead to more stable feature selection algorithms. There are various strategies for measuring stability based on index and rank, but only a few techniques for measuring stability based on weights exist. Only Pearson's correlation coefficient, among the numerous stability measures, computes the stability by taking the feature's weight into account. Because the high-dimensional dataset contains highly associated features, group feature selection was performed. A suggested algorithm's used SVM experimental findings show that it has the highest accuracy (an average of 91.1%) when compared to similar approaches. In addition, each dataset has an average of 7.12 unique features, resulting in excellent efficiency and the achievement of a subset of compressed features.In terms of classification methods, it can be seen that SVM was the most widely used method for both AD and MCI categorization.

VIII. CONCLUSIONS

The importance of feature correlation in feature selection should not be underestimated. The differential Pearson's Correlation Coefficient information entropy was used to suggest a new feature selection approach, also produced was an algorithm framework based on differential correlation information entropy, to increase classification performance by taking full use of the highest possible correlation between features. There are drawbacks to several feature selection strategies, and the suggested solution addresses them. As a result, the

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

suggested method correctly recognizes feature relationships and eliminates redundant and irrelevant features of the selected feature set. The methods were applied to the feature subsets and then their performances were compared to reveal which approach is better for early AD detection. This research found that using the SVM method with a feature set generated via a heterogeneous approach yielded the best classification accuracy.

Our future study will focus on the application of heterogeneous methodologies to increase classification accuracy by combining diverse data mining methods.

REFERENCES

[1] L. Wang and Z. P. Liu, "Detecting diagnostic biomarkers of Alzheimer's disease by integrating gene expression data in six brain regions," *Frontiers in Genetics*, vol. 10, no. MAR, pp. 1–11, 2019, doi: 10.3389/fgene.2019.00157.

[2] 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 International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings*, pp. 1–6, 2020, doi: 10.1109/MERCon50084.2020.9185336.

[3] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks," *Advances in Bioinformatics*, vol. 2015, pp. 1–9, 2015, doi: 10.1155/2015/639367.

[4] 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 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020*, pp. 15–20, 2020, doi: 10.1109/ICDABI51230.2020.9325640.

[5] S. J. Mohammed, "A Proposed Alzheimer ' s Disease Diagnosing System Based on Clustering and Segmentation Techniques," *Engineering and Technology*, vol. 36, no. 2, pp. 160–165, 2018.

[6] A. T. Hashim and G. S. Mahdi, "Modification of RC5 Algorithm for Image Encryption 1," *Iraqi J. Comput. Commun. Control Syst. Eng*, vol. 14, no. 2, 2014.

[7] K. Nishiwaki, K. Kanamori, and H. Ohwada, "Gene Selection from Microarray Data for Alzheimer's Disease Using Random Forest," *International Journal of Software Science and Computational Intelligence*, vol. 9, no. 2, pp. 14–30, 2017, doi: 10.4018/ijssci.2017040102.

[8] M. M. Mahmoud and A. R. Nasser, "Dual Architecture Deep Learning Based Object Detection System for Autonomous Driving," *Iraqi J. Comput. Commun. Control Syst. Eng*, vol. 21, no. 2, pp. 36–43, 2021.

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

[10] N. Jameel and H. S. Abdullah, "Intelligent Feature Selection Methods : A Survey," *Engineering and Technology Journal*, vol. 39, no. 01, pp. 175–183, 2021.

[11] S. Selvaraj and J. Natarajan, "Microarray Data Analysis and Mining Tools," vol. 6, no. 3, 2011.

[12] A. A. Abdulhussein and F. A. Raheem, "Hand Gesture Recognition of Static Letters American Sign Language (ASL) Using Deep Learning," *Engineering and Technology Journal*, vol. 38, no. 6A, pp. 926–937, 2020, doi: 10.30684/etj.v38i6a.533.

[13] 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," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 16, p. 95, 2021, doi: 10.3991/ijim.v15i16.24191.

[14] 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," *Journal of Advanced Computer Science and Technology Research*, no. December, 2016.

[15] A. T. Sadiq and K. S. Musawi, "Modify Random Forest Algorithm Using Hybrid Feature Selection Method," *International Journal on Perceptive and Cognitive Computing*, vol. 4, no. 2, pp. 1–6, 2018, doi: 10.31436/ijpcc.v4i2.59.

[16] H. N. Abdullah, "Deep CNN Based Skin Lesion Image Denoising and Segmentation using Active Contour Method," *Engineering and Technology Journal*, vol. 37, no. 11, 2019.

[17] I. Jo, S. Lee, and S. Oh, "Improved measures of redundancy and relevance for mRMR feature selection," *Computers*, vol. 8, no. 2, pp. 1–14, 2019, doi: 10.3390/computers8020042.

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

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

[19] 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, doi: 10.1186/1471-2105-13-266.

[20] K. Lunnon *et al.*, "A blood gene expression marker of early Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 33, no. 3, pp. 737–753, 2013, doi: 10.3233/JAD-2012-121363.

[21] 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, doi: 10.1186/1471-2105-15-S16-S11.

[22] S. Sood *et al.*, "A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status," *Genome Biology*, vol. 16, no. 1, pp. 1–17, 2015, doi: 10.1186/s13059-015-0750-x.

[23] 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 - Proceedings of the 23rd Iranian Conference on Electrical Engineering*, vol. 10, pp. 956–959, 2015, doi: 10.1109/IranianCEE.2015.7146349.

[24] S. Zahra Paylakhi, S. Ozgoli, and S. Paylakhi, "Identification of Alzheimer disease-relevant genes using a novel hybrid method," *Progress in Biological Sciences*, vol. 6, no. 1, pp. 37–46, 2016, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed

[25] N. Voyle *et al.*, "A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis," *Journal of Alzheimer's Disease*, vol. 49, no. 3, pp. 659–669, 2015, doi: 10.3233/JAD-150440.

[26] M. Barati and M. Ebrahimi, "Identification of Genes Involved in the Early Stages of Alzheimer Disease Using a Neural Network Algorithm," *Gene, Cell and Tissue*, vol. 3, no. 3, 2016, doi: 10.17795/gct-38415.

[27] M. Balamurugan, A. Nancy, and S. Vijaykumar, "Alzheimer's disease diagnosis by using dimensionality reduction based on KNN Classifier," *Biomedical and Pharmacology Journal*, vol. 10, no. 4, pp. 1823–1830, 2017, doi: 10.13005/bpj/1299.

[28] H. Li *et al.*, "Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017, doi: 10.1038/s41598-017-13700-w.

[29] 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, doi: 10.3390/molecules23123140.

[30] Inzhong *et al.*, "Systematic Analysis and Biomarker Study for Alzheimer's Disease," *Scientific Reports*, vol. 8, no. 1, pp. 1–14, 2018, doi: 10.1038/s41598-018-35789-3.

[31] K. Sekaran and M. Sudha, "Diagnostic gene biomarker selection for alzheimer's classification using machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 2348–2352, 2019, doi: 10.35940/ijitee.L3372.1081219.

[32] T. Lee and H. Lee, "Prediction of Alzheimer's disease using blood gene expression data," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-60595-1.

[33] K. P. Muhammed Niyas and P. Thiyagarajan, "Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification," *Journal of King Saud University - Computer and Information Sciences*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2020.12.009.

[34] R. A. Saputra *et al.*, "Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization," *Journal of Physics: Conference Series*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012025.

[35] 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 Systems with Applications*, vol. 140, p. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.

[36] 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," *Medicine in Omics*, vol. 1, no. December 2020, p. 100001, 2021, doi: 10.1016/j.meomic.2020.100001.

[37] C. Park, J. R. Kim, J. Kim, and S. Park, "Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles," *PLoS ONE*, vol. 13, no. 7, pp. 1–15, 2018, doi: 10.1371/journal.pone.0201056.

[38] A. K. Hassan and S. N. Mohammed, "A novel facial emotion recognition scheme based on graph mining," *Defence Technology*, vol. 16, no. 5, pp. 1062–1072, 2020, doi: 10.1016/j.dt.2019.12.006.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

[39] Z. Manbari, F. AkhlaghianTab, and C. Salavati, "Hybrid fast unsupervised feature selection for high-dimensional data," *Expert Systems with Applications*, vol. 124, pp. 97–118, 2019, doi: 10.1016/j.eswa.2019.01.016.

[40] H. A. R. Akkar and S. A. Salman, "Detection of Biomedical Images by Using Bio-inspired Artificial Intelligent," *Engineering and Technology Journal*, vol. 38, no. 2A, pp. 255–264, 2020, doi: 10.30684/etj.v38i2a.319.

[41] M. S. and Asst. Prof., "Handwriting Word Recognition Based on SVM Classifier," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, pp. 64–68, 2015, doi: 10.14569/ijacsa.2015.061109.

[42] O. D. Madeeh and H. S. Abdullah, "An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market," *Journal of Physics: Conference Series*, vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012008.

[43] X. Huang *et al.*, "Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning," *BMC Neurology*, vol. 18, no. 1, pp. 1–8, 2018, doi: 10.1186/s12883-017-1010-3.

[44] W. M. S. Abedi, I. Nadher, and A. T. Sadiq, "Modified deep learning method for body postures recognition," *International Journal of Advanced Science and Technology*, vol. 29, no. 2, pp. 3830–3841, 2020.

[45] B. Venkatesh and J. Anuradha, "A review of Feature Selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/CAIT-2019-0001.

[46] Z. A. Mohammed, M. N. Abdullah, and I. H. Al-hussaini, "Predicting Incident Duration Based on Machine Learning Methods," *Iraqi Journal of Computer, Communication, Control and System Engineering*, pp. 1–15, 2021, doi: 10.33103/uot.ijccce.21.1.1.

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

[48] 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.

[49] M. Salam and A. A. Hassan, "Offline isolated arabic handwriting character recognition system based on SVM," *International Arab Journal of Information Technology*, vol. 16, no. 3, pp. 467–472, 2019.

[50] A. R. T. Silva *et al.*, "Transcriptional Alterations Related to Neuropathology and Clinical Manifestation of Alzheimer's Disease," *PLoS ONE*, vol. 7, no. 11, 2012, doi: 10.1371/journal.pone.0048751.

[51] A. K. Tiwari, "Machine Learning Based Approaches for Prediction of Parkinson's Disease," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 33–39, 2016, doi: 10.5121/mlaij.2016.3203.

[52] K. Qu, F. Gao, F. Guo, and Q. Zou, "Taxonomy dimension reduction for colorectal cancer prediction," *Computational Biology and Chemistry*, vol. 83, no. June, p. 107160, 2019, doi: 10.1016/j.compbiolchem.2019.107160.

[53] M. F. Demiral, "A parameters analysis of sine cosine algorithm on travelling salesman problem," *El-Cezeri Journal of Science and Engineering*, vol. 7, no. 2, pp. 526–535, 2020, doi: 10.31202/ecjse.662864.

[54] Y. Zhang, Y. Zhou, D. Zhang, and W. Song, "A stroke risk detection: Improving hybrid feature selection method," *Journal of Medical Internet Research*, vol. 21, no. 4, 2019, doi: 10.2196/12437.

[55] H. R. Yassein, N. M. G. Al-Saidi, and A. K. Farhan, "A new NTRU cryptosystem outperforms three highly secured NTRU-analog systems through an innovational algebraic structure," *Journal of Discrete Mathematical Sciences and Cryptography*, no. May, 2020, doi: 10.1080/09720529.2020.1741218.

[56] S. Mahajan, G. Bangar, and N. Kulkarni, "Machine Learning Algorithms for Classification of Various Stages of Alzheimer ' s Disease: A review," *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. 08, pp. 817–824, 2020.

[57] 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, doi: 10.1371/journal.pone.0150624.

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

[59] I. S. Abed, "Lung Cancer Detection from X-ray images by combined Backpropagation Neural Network and PCA," *Engineering and Technology Journal*, vol. 37, no. 05, 2019.

[60] S. Geman, E. Bienenstock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992, doi: 10.1162/neco.1992.4.1.1.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.22.4.2

[61] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics*, vol. 112, no. 2, pp. 1173–1181, 2020, doi: 10.1016/j.ygeno.2019.07.002.

[62] P. Zheng *et al.*, "Crystallographic Analysis of the Catalytic Mechanism of Phosphopantothenoylcysteine Synthetase from Saccharomyces cerevisiae," *Journal of Molecular Biology*, vol. 431, no. 4, pp. 764–776, 2019, doi: 10.1016/j.jmb.2019.01.012.

[63] L. Billeci, A. Badolato, L. Bachi, and A. Tonacci, "Machine learning for the classification of alzheimer's disease and its prodromal stage using brain diffusion tensor imaging data: A systematic review," *Processes*, vol. 8, no. 9, 2020, doi: 10.3390/pr8091071.

[64] L. Mesrob *et al.*, "Identification of atrophy patterns in Alzheimer's disease based on SVM feature selection and anatomical parcellation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5128 LNCS, pp. 124–132, 2008, doi: 10.1007/978-3-540-79982-5_14.

[65] J.Neelaveni and M.S.Geetha Devasana, "Alzheimer disease predication using machine learning," *IEEE*, vol. 978-1-7281–5197, pp. 101–104, 2020.

[66] N. Jameel and H. S. Abdullah, "A Proposed Intelligent Features Selection Method Using Meerkat Clan Algorithm," *Journal of Physics: Conference Series*, vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012061.

Received 29/November/2021; Accepted 16/February/2022
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*