# Machine Learning-based ALS Diagnosis Using Gene Expression Data

Duc-Long Vu
*Department of Data engineering*
*Posts and Telecommunications Institute of Technology*
Hanoi, Vietnam
longvd@ptit.edu.vn

Hai-Chau Le
*Department of Data engineering*
*Posts and Telecommunications Institute of Technology*
Hanoi, Vietnam
chaulh@ptit.edu.vn

***Abstract*—Amyotrophic Lateral Sclerosis (ALS) is a rare disease that debilitates the body of a patient with few treatment methods. The biological insight involved in ALS remains elusive and faces challenges in making decisions about diagnosis. Recently, gene expression has been valuable for overcoming obstacles in analysis and providing accurate diagnosis outcomes in many diseases. In this work, we presented a new method concentrating on biomarker selections including three different procedures of gene reduction. We retrieved datasets and conducted differential expression analysis to identify the gene markers that vary significantly. We then apply filter and embedded methodology for gene selection with the Maximum Relevance Minimum Redundancy (MRMR) followed by the least absolute shrinkage and selection operator (LASSO) regression. Various machine learning algorithms corresponding with a set of gene combinations are then estimated using a series of cross-validation procedures. The optimal gene subset corresponding with the machine learning algorithms is then validated separately on the testing dataset and considered the best model based on the receiver operating characteristic (ROC) curve. Finally, a combination including 22 potential genes with Logistic Regression algorithms is considered the most effective method for diagnosis of ALS with an AUC score of 87.90% which is dominant in comparison with other current methods.**

***Index Terms*—Machine Learning, Feature Selection, Gene Expression, Amyotrophic Lateral Sclerosis, Biomarkers.**

## I. Introduction

Amyotrophic lateral sclerosis (ALS) is considered a rare disease that tremendously affects primary neurodegenerative and leads to dysfunction of the human motor system. According to the recording in Europe, among 100,000 individuals, the number of people dealing with ALS varies between 2 and 3 [1] and the prevalence is between 5 and 8 in 100,000 cases [2]. The ALS Therapy Development Institute also proposes that around 450,000 people in the world are affected by ALS. Although a common underlying cause hasn't been identified for all variations of ALS, it is expected that the frequency of the condition will increase over the next few decades leading to the emergence of efficient methods for the discovery of drugs and treatment methods [3]. Besides, due to the inconsistency of the clinical and genetics of ALS, there are considerable differences between the liability profile, hence flexible individualized care strategies are required in clinical treatment. Additionally, the requirement for personalized pharmaceutical therapies would also be considered [4].

While a definitive cure for ALS remains elusive, managing its symptoms can substantially enhance the quality of the patient and extend the survival for those impacted by the condition [5]. Due to the diagnosis of ALS occurring with a considerable delay, many patients miss the opportunity for early treatment which affects a lot to the development of this disease [6]. In the study [7], the author reveals that replenishment of NAD+ can significantly improve the clinical features of patients with ALS and lead to a potential method for treatment in ALS [8]. These reasons underscore the life-saving potential of effective methods and tools for accurately predicting the prevalence and occurrence of ALS.

In recent years, the world witnessed an explosion with the increasing of Machine Learning (ML) algorithms, and a large number of research work has been done using ML for diagnosis and prognosis of the condition of ALS. However, the current works are usually only applied to small clinical data, power calculation, or statistical estimation which do not reveal a lot of information about ALS and give reliable results due to ALS consisting of massive biological function disorders. With motivation for constructing an effective model for the diagnosis of ALS, many scientists have been applying machine learning, and deep learning techniques to address this obstacle. In [9], the author proposed a method using biomedical images with signal processing techniques to identify the imaging features that are related to ALS. The feature selection technique followed by a cross-validation procedure using the Logistic Regression method is also considered to provide acceptable results. Moreover, in [10], the authors proposed a noble method using combined image metrics for developing a diagnosis model that the model performance relies only upon clinical features [11], [12] with an average score in accuracy reaching 90%. However, the recent model faces limited data size, and a lower true positive rate but dramatically increases the false negative rates. This issue could be explained by the unbalanced of the samples in the ALS dataset as a rare disease. Besides, the more practicals need to be studied to provide more information to develop a robust model of diagnosis and leverage the treatment of ALS.

Moreover, ALS is also observed in approximately 5-10 % of cases that have a family history of the disease. The majority with around 90% of ALS patients experience in condition

without any family history. In [13], [14], the author reveals 30 genes contributed significantly to ALS, and some mutation genes are also associated with ALS. Despite the knowledge of ALS, the trustworthy causes and signature mechanisms of this disease are a mystery. A vast of the current effort in analyzing the genomic data is to identify the most effective biomarkers related to ALS and capable of explaining the pathogenic mechanism of this disease. Nevertheless, despite the development of genomic technology and the variety of genomic datasets, the current research on this area for ALS disease has witnessed a shortcoming. Leveraging the efficiency of ML and DL, in [15], the author proposed a noble method in a proposed noble method using Capsule Network and Principal Component Analysis (PCA) for developing a diagnosis model for ALS based on individual genotype profiles with remarkable results. Additionally, relying on the gene expression data of ALS patients, combined with the WGCNA and LASSO regression algorithm, the author of [16] identifies a combination of five genes that is significant to the developed diagnosis model for ALS. In [17], 850 genes and 468 principle components were identified as the potential biomarker for ALS diagnosis using S algorithm. Moreover, using several classification models, the expression-based model from [18], does not provide sufficient results in discriminating between ALS with ALS-mimic disease. Therefore, it can be inferred that the utilization of blood gene expression markers for predicting ALS patients yielded insufficient results. Besides, all the studies above also require a complex method for identifying ALS biomarkers and constructing a potential diagnosis model. These results raise questions about revealing the biological insight from the gene expression profile, and how to perform an effective model for identifying the important biomarkers for ALS, but with less effort in the computational aspect.

Motivated by the obstacles above, in this work, we proposed a novel approach based on data analysis to select the optimal genes subset which can contribute significantly to constructing the diagnosis model with less complexity but maintain the overall performance. We propose a sequential gene selection procedure that leverages the advantages of three feature selection algorithms to point out the optimal set of genes that can be used to create an effective model for diagnosing ALS. Additionally, we use the two largest datasets considered the best resource available for identifying ALS biomarkers from the peripheral blood of ALS patients and control for this research.

## II. Data and Preprocessing

### A. Data source

Since the limitation of the source of the gene expression publication dataset for ALS disease, in our paper, we use only the dataset GSE112681 The dataset GSE112681 comprises information from datasets GSE112676 and GSE112680, both obtained from the GEO database [18]. These datasets originate from gene expression profiling of whole blood. The data is gathered using two distinct platforms: Illumina HumanHT-12 V3.0 and HumanHT-12 V4.0 expression Bead-Chip arrays. The expression values of two datasets are then extracted by applying the *GEOquery* R package. Besides, the information of samples is extracted from the *series matrix* file. There are a total of 1042 samples including both ALS (397 patients) and normal control (645 people) in two datasets. It is noteworthy to inform that, in our work we only considered the analysis in gene expression of control and ALS patients, so 75 ALS-mimic diseases in GSE112680 are excluded.

### B. Filtering of probes and data preprocessing

In both datasets, we collect and analyze the data from the raw data downloaded from the GEO database. The probes in raw data that met all the following criteria were eliminated: (1) the probes and those with no specific symbol; (2) the probes not corresponding to the genes symbol. The microarray file annotation file appropriate for each platform was used for mapping the probes with corresponding gene symbols. The probes associated with multiple gene symbols were removed while the average values of the expression values were calculated for genes that corresponded to multiple probes. The gene expression data with appropriate gene symbols is then considered as the input values for further analysis. It is noteworthy that, due to the differences in the representation platform of the two datasets, only the gene symbols included in both datasets are considered in this work. The expression data in two datasets with annotation were combined for the following step.

We used the *limma* R package to determine differential expression genes with Benjamini-Hochberg (BH) as the correction method. The fold-change value relative to ALS indicates that the up-regulated genes are those whose expression is higher in the ALS samples compared to the Depression of the same genes in the CON samples. The expression values of selected genes are based on the fold-change and the *p*-value then rescaling into the range 0 to 1 using the *min-max* normalization algorithm. Then, combined data is divided into two parts named training and testing data with ratios of 80% and 20%, respectively. We also note that the proportion of the ALS samples and Control samples in two partitions are the same.

## III. Methodology

In this work, our proposed method is shown in Fig.1 includes three main stages. Firstly, Data preprocessing is implemented for removing the outlier data, and cleaning the gene expression data becomes more efficient for further procedure. Here, the number of the gene markers drops significantly by the differentially expressed analysis process. Secondly, two algorithms of feature selection are performed to point out the most informative genes that are important for the diagnosis of ALS. Then, the construction of different gene subsets was produced by the Sequential Forward Feature Selection (SFFS) algorithm based on the selected genes. There is a 5-fold cross-validation procedure then performed for each machine learning algorithm appropriate with the gene combination to reveal the best gene combinations and machine learning model

for diagnosis of ALS. Lastly, the proposed gene combination is validated on the testing set with several machine-learning models in the Model Estimation phase. The gene subset which is adopted as the input of a classifier producing the highest validation diagnosis performance in terms of AUC score, is chosen as the proposed gene subset and algorithm for ALS diagnosis.

We use five machine learning algorithms in this work, which are K-nearest neighbors (KNN), Extreme Gradient Boosting (XGB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) [19]. These machine-learning algorithms are used to estimate the efficiency of the proposed gene combination as well as play a key role in developing a diagnosis model.

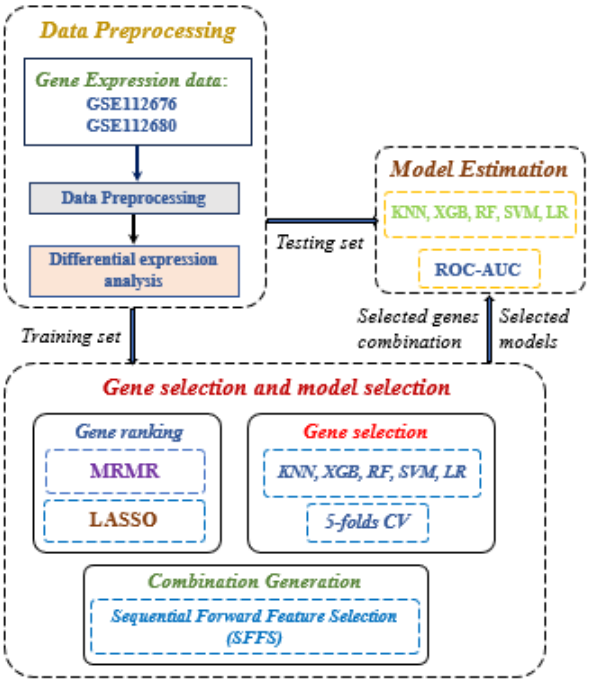


Fig. 1. Flow method

### A. Gene Importance Analysis

We introduce a novel approach to analyzing the importance of parameters in constructing a sufficient and accurate diagnosis model for ALS. Motivated by the efficiency of the feature selection method in extracting the important features (in this scenario the genes) for deploying the machine learning model, a sequential filter, embedded, and wrapper methods are considered to produce the most value genes which are significantly related to ALS. Firstly, we use the MRMR method to evaluate the correlation between the differential expressed genes and the relevance between individuals with the target variable. Then, the LASSO regression algorithm is used to calculate the coefficient values of the genes for reducing the redundant features. We briefly explain the two algorithms in the following section.

*1) Maximum Relevance Minimum Redundancy (MRMR):* The MRMR method is a feature selection technique designed to choose a subset of features that optimizes the relationship with the target variable, while also minimizing redundancy among the selected features. There are two main parts of MRMR Relevance (R) and Redundancy (D).

Assuming there are a total of *S* features in the set of features $X_i(i \in 1, 2, ..., S)$. Its feature importance $(M_i)$ based on the MRMR criterion can be expressed as:

$$M(i, S) = R(i) - \alpha D(i, S) \tag{1}$$

with $\alpha$ as the weighting parameter that the balance trade-off between relevance and redundancy.

*2) Lasso Logistic Analysis:* The LASSO algorithm is a linear model that combines logistic regression with L1 regularization. It models the relationship between the features and the binary target variable using the logistic function. With the L1 regularization, the objective function of the model encourages small coefficients and consequently, sparsity in the model. In mathematic formulas, the L1 regularization is the sum of the absolute values of the coefficients:

$$L_1 = \lambda \sum_{j=1}^{p} |w_j| \tag{2}$$

with $\lambda$ as the regularization parameter that controls the regularization strength. Higher $\lambda$ leads to more coefficients being become zero; $p$ is the number of features.

The goal is to find the coefficient values $w$ that minimize the LASSO objective function. This can be done using the optimization technique, in this case, gradient-descent methods. The L1 regularization encourages some feature coefficients to become zero, leading to feature selection. Features with the coefficient is zeros are effectively excluded from the model, resulting more interpretable and potentially simpler model.

### B. Gene Combination Construction

The remaining genes after two phases of feature selection are considered the most informative genes for developing a machine learning model. Based on the absolute coefficient value, we rank the genes from highest to lowest and produce the ranking table. Following by applying the Sequential Forward Feature selection (SFFS) procedure, we constructed the gene combination from the ranking table. Particularly, an ordered list of genes is formulated, encompassing diverse gene permutations. The initial permutation comprises the gene situated at the peak of the ranked list, characterized by the utmost significance value. Subsequently, the subsequent permutations encompass pairs of genes, with the first one being the highest-ranked gene and the second one occupying the second position in the list. This progression is outlined in Algorithm 1, detailing the approach for generating distinct gene combinations.

### C. Diagnosis Model Development

All the gene combinations are then fed into different machine-learning models to identify the optimal learning and

**Algorithm 1** Gene Combination Generation

*(1) Ranking Genes based on absolute coefficient score*
*(2) Gene combination generation:*
*- Number of potential gene: n*
- Combine of genes subset: S = [];
- Initial number of gene in subset: $g$ = 1

**Repeat**

- Construct genes subset based on a gene coefficient table from 1 to $g$;
- Append the gene subset to $S$; $g = g + 1$

**Until** $g = n$;

structure parameters on the training set. There are five machine learning models constructed with $n$ combinations. Hence, there are a total of 5 x n models considered. Hyper-parameter tuning is a requirement for each classifier with an appropriate gene subset to overcome the overfitting problems. Additionally, the grid search with $k$ folds cross-validation process is considered in this stage to address the optimal parameter for all models. Furthermore, the optimal gene subset corresponding with the classifier that produces the best result is then evaluated performance on the testing set.

### D. Model Evaluation

The optimal gene subset, which is the output of the model development phase, is then estimated for the diagnosis performance on the testing set. Different ML algorithms are then trained and tested on the training set and testing set using the optimal combinations. These models utilize the identified gene combinations as inputs for accurately diagnosing ALS. The algorithms for diagnosis are chosen based on the superior performance exhibited by the corresponding machine learning models.

## IV. Experiments and Results

We use a number of evaluation metrics including accuracy (Acc), sensitivity (Se), specificity (Sp), and AUC-ROC score (AUC) to estimate the results of our ML model. The AUC score and Accuracy are the main metrics we use to measure the performance of the optimized model. Sn and Sp were used to evaluate the effectiveness of the model in addressing the ALS and control samples correctly.

### A. Differential Expressed Gene Analysis

The gene expression and annotation data were retrieved and then normalized using the gcRMA method in *R* package. After, the probe filtering and gene symbol mapping process, a huge number of gene markers and probes in both data sets were eliminated. In the last filtering step, GSE112676, and GSE112680 datasets included 16878 and 16837 genes, respectively. However, as we mentioned before, only genes that appeared in both datasets are considered for further analysis hence, finally 16833 genes are identified in both gene expression sets.

The combination of two datasets is then normalized and fed into a differential expressed analysis procedure. Surprisingly, by setting the threshold for the absolute fold change $\geq 1.5$ and adjusted $p$-value $\leq 0.05$, only 33 genes from a total of 16833 genes were found to be DEGs between ALS and control samples. These genes are ranked based on the absolute fold change value and then fed into the following gene selection procedure.

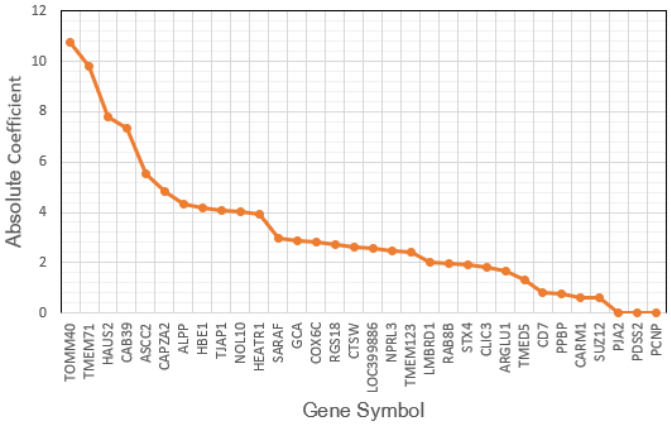


Fig. 2. Genes LASSO Absolute Coefficient

### B. Gene ranking

From 33 genes DEGs received from the differential expression analysis, two layers of feature selection algorithms are applied to obtain the final gene ranking table.

*1) MRMR score:* We implement the MRMR feature selection algorithm by using the Scikit-learn library using Python. In our work, the mutual information between the feature are also calculated based on this library to receive the Redundancy and Relevance value. From 33 genes, the MRMR score of each feature was obtained and ranked from the lowest to the highest. The gene symbol YPEL5 shows the lowest MRMR score with results approximate zeros, so they are eliminated. After this first filter process, only 32 genes are remaining.

*2) LASSO coefficient:* The combination including 32 DEGs then fed into the LASSO logistic model to estimate the coefficient values for individual genes. The hyper-parameter tuning for this model is necessary with hyper-parameter C and $\lambda$. The 5-fold cross-validation procedure was also applied. The coefficients for each feature in the logistic model with the L1-regularization are illustrated in Fig.2. Clearly, three genes PJA2, PDSS2, and PCNP show the coefficient approximate zeros, in other words, these genes are considered non-informative or less important for classification tasks, so we ignore these genes. Consequently, after two phases of feature elimination, there are 29 genes remain the most important features for developing the ALS diagnosis model. Based on the absolute coefficient, we ranked the gene symbol from the highest to the lowest to construct a gene ranking table for further experiments.

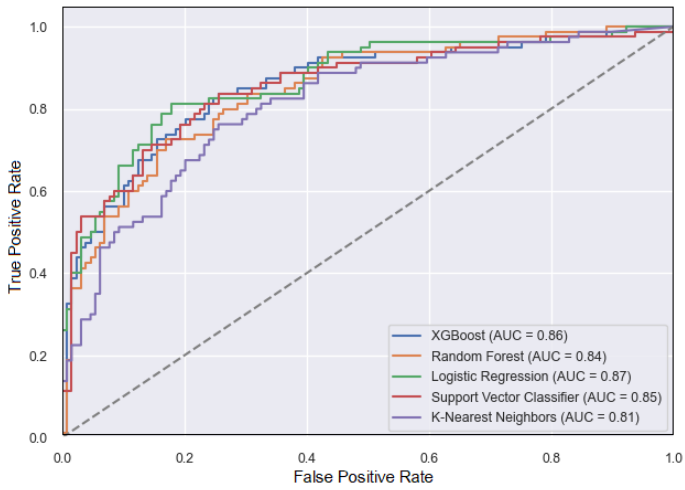


Fig. 3. AUC score of five ML algorithms on the testing set with optimal gene set

### C. Diagnosis Model Development

The outcomes of the gene ranking process determine the optimal number of *m* genes to be selected. This is achieved by utilizing a chosen classifier and supplying it with a subset of ranked genes. The process starts with the highest-ranked gene and proceeds downwards, ultimately selecting *m* genes that yield the highest classification score. Based on Algorithm.1, there are a total of 29 gene combinations that are estimated to figure out the most effective gene group. A total of 5 ML models and 29 gene subsets are considered in this work. Hence, there are a total of 145 models constructed to search for the best gene combination corresponding with the machine learning model. These classifiers estimated their diagnosis efficiency on the training cohorts with a ten times five-fold cross-validation procedure. Table. I shows the results of five models, which produce the highest AUC score corresponding to each associated with the highest classification accuracy achieved through the optimal gene combinations. It is worth highlighting that the selection of each model in Table. I was made after comparing classification accuracies across 145 models employing gene combinations. Notably, the mean values align with the most efficient gene combination identified by the Logistic Regression (LR) classifier. As a result, the selection of these 22 genes as the most informative gene subsets relevant to ALS is substantiated.

TABLE I
MODEL VALIDATION RESULTS

| Model | #genes | Acc (%) | Sn (%) | Sp (%) | AUC (%) |
|---|---|---|---|---|---|
| ***LR*** | ***22*** | ***82.28 ± 3.80*** | ***72.86 ± 5.46*** | ***88.94 ± 3.24*** | ***87.90 ± 4.37*** |
| KNN | 8 | 77.18 ± 2.20 | 62.43 ± 5.64 | 86.2 ± 4.37 | 84.25 ± 2.65 |
| SVM | 27 | 81.86 ± 2.92 | 71.91 ± 5.09 | 87.98 ± 2.58 | 87.72 ± 4.20 |
| RF | 11 | 80.18 ± 3.69 | 69.06 ± 6.79 | 87.00 ± 3.00 | 84.94 ± 4.06 |
| XGB | 25 | 82.95 ± 1.62 | 74.75 ± 3.40 | 87.98 ± 1.89 | 86.69 ± 1.77 |

### D. Model Validation

To demonstrate the effectiveness of the optimal gene subset, we evaluate the performance of the diagnosis model on the testing set. Particularly, five ML model was trained on the training set and then estimated precisely on the testing set separately. It is noteworthy to clear that all five machine learning models are also optimized with a grid-search method nested by a five-fold cross-validation procedure for searching optimal hyper-parameters for 22 gene combinations. Table.II shows the detail of the ML classifier on the testing set. The diagnosis performance of the LR model is still the highest compared with another ML algorithm in all evaluation metrics. Additionally, the ROC-AUC curve is also illustrated in Fig.3 with the AUC score being round with two numbers. Moreover, we also compare the results of our work with the two latest studies which to our best knowledge given state-of-the-art results using similar ALS genome datasets in Table. III. As a result, we suggest the combination of 22 ALS-related genes corresponding to the LR model as the most effective method for diagnosis of ALS disease - one of the most complicated and lacks research in the genome which is considered as a signature biomarker of ALS.

TABLE II
MODEL ESTIMATION RESULTS ON TESTING SET

| Model | Acc (%) | Sn (%) | Sp (%) | AUC (%) |
|---|---|---|---|---|
| ***LR*** | ***81.81*** | ***76.25*** | ***85.27*** | ***86.75*** |
| KNN | 74.16 | 62.50 | 81.39 | 81.06 |
| SVM | 79.42 | 71.25 | 84.49 | 85.29 |
| RF | 78.46 | 70.00 | 83.72 | 84.48 |
| XGB | 79.42 | 72.50 | 83.72 | 85.62 |

TABLE III
COMPARISON OF THE PROPOSED ALGORITHM TO EXISTING WORKS

| Study | Acc (%) | Sn (%) | Sp (%) | AUC (%) |
|---|---|---|---|---|
| Proposed algorithm | 82.28 | 72.86 | **88.94** | **87.90** |
| [16] | - | - | - | 86.5 |
| [17] | **87.00** | **86.00** | 87.00 | - |

## V. DISCUSSION

Our proposed ML method can identify the most informative gene markers that are significantly related to ALS disease. The efficient diagnosis model is developed using Logistic Regression algorithms and our work addresses the complicated processing and accuracy problem of the ALS diagnosis models. A potent combination of 22 genes has been chosen as the ultimate gene subset to perform remarkable results in the diagnosis of ALS. Our proposed gene subset is significantly small, compared with the existing work like in [17] with 850 genes, and our model with the selected gene combination provides better performance compared to the model given in [16]. The reason behind the successful identification of such a gene combination is the utility of the three procedures of gene reduction that summarize the pros

of filter, embedded, and wrapper feature selection algorithms. Moreover, the method of analyzing differentially expressed genes has gained widespread usage for the examination of gene expression profiles. By applying this method as the first gene elimination procedure, only 33 genes were chosen from 16833 genes which contributed significantly to analyzing the disease. Based on the differential expressed gene subsets, the MRMR algorithms interpret the relationship between the genes with the outcome of the genome profile as well as the correlation between individual genes. The lower score indicates the less important genes, so we can eliminate the genes that are not sufficient for ALS diagnosis. Additionally, the LASSO LR model was also considered in our work for the calculation of the coefficient of the logistic model in predicting the disease. These values are assigned for each feature as the linear combination with the input data (i.e. gene expression data) for the objective function in developing the classifier. Intuitively, the closer to zeros, the more the appropriate genes to this coefficient are less efficient for the diagnosis model. Consequently, combining two feature selection approaches can significantly reduce unimportance genes and retain the most informative genes which are useful for developing a diagnosis model.

On the other hand, the SFFS algorithm is applied for searching the optimal gene combination from the selected genes for defining the classifier. To develop an innovative algorithm for diagnosing ALS, various ML models have been employed to assess the diagnostic efficacy of the chosen gene combination. The simulation results indicate that our proposed model using LR with the optimal subset of 22 genes outperforms the existing works on ALS diagnosis using the same genome dataset. Moreover, a compact set of genes, identified using the gene selection procedure outlined, serves to emphasize the viability and efficiency of the suggested algorithm for potential clinical implementations.

## VI. Conclusion

We proposed an efficient approach using machine learning combined with gene analysis to develop a robust model for the diagnosis of ALS. The combination includes 22 genes determined using the sequential of three gene selection algorithms. We leverage the efficiency of MRMR and LASSO algorithms in the feature selection procedure to reveal the most important genes to construct a diagnosis model. Then, a sequential forward feature selection method is applied to search for an optimal gene combination that is suitable for the machine learning algorithm. The selected 22 genes combination combined with the LR algorithms show significant results in diagnostics and robustness with an average AUC score of 87.9%. In addition, the effectiveness of our method as well as the intelligent diagnosis model also demonstrated as superior to the result of some novel studies. Our research phase proposed a robust method in gene selection and constructed a precise predictive model for ALS disease, which has the potential to be utilized as a foundation for both clinical diagnostic testing and in-depth biological mechanistic investigations.

## References


[1] O. Hardiman, A. Al-Chalabi, A. Chio, E. M. Corr, G. Logroscino, W. Robberecht, P. J. Shaw, Z. Simmons, and L. H. van den Berg, "Amyotrophic lateral sclerosis," *Nature Reviews Disease Primers*, vol. 3, p. 17071, 10 2017.

[2] A. Chiò, G. Logroscino, B. Traynor, J. Collins, J. Simeone, L. Goldstein, and L. White, "Global epidemiology of amyotrophic lateral sclerosis: A systematic review of the published literature," *Neuroepidemiology*, vol. 41, pp. 118–130, 2013.

[3] K. C. Arthur, A. Calvo, T. R. Price, J. T. Geiger, A. Chiò, and B. J. Traynor, "Projected increase in amyotrophic lateral sclerosis from 2015 to 2040," *Nature Communications*, vol. 7, p. 12408, 8 2016.

[4] J. P. V. den Berg, S. Kalmijn, E. Lindeman, J. H. Veldink, M. de Visser, M. M. V. der Graaff, J. Wokke, and L. H. V. den Berg, "Multidisciplinary als care improves quality of life in patients with als," *Neurology*, vol. 65, pp. 1264–1267, 10 2005.

[5] R. G. Miller, C. E. Jackson, E. J. Kasarskis, and et al., "Practice parameter update: The care of the patient with amyotrophic lateral sclerosis: Drug, nutritional, and respiratory therapies (an evidence-based review): Report of the quality standards subcommittee of the american academy of neurology," *Neurology*, vol. 73, pp. 1218–1226, 10 2009.

[6] R. H. Brown and A. Al-Chalabi, "Amyotrophic lateral sclerosis," *New England Journal of Medicine*, vol. 377, pp. 162–172, 7 2017.

[7] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, pp. 942–955, 3 2011.

[8] S. Lautrup, D. A. Sinclair, M. P. Mattson, and E. F. Fang, "Nad+ in brain aging and neurodegenerative disorders," *Cell Metabolism*, vol. 30, pp. 630–655, 10 2019.

[9] J. E. de la Rubia, E. Drehmer, J. L. Platero, and et al., "Efficacy and tolerability of eh301 for amyotrophic lateral sclerosis: a randomized, double-blind, placebo-controlled human pilot study," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 20, pp. 115–122, 1 2019.

[10] C. Schuster, O. Hardiman, and P. Bede, "Development of an automated mri-based diagnostic protocol for amyotrophic lateral sclerosis using disease-specific pathognomonic features: A quantitative disease-state classification study," *PLOS ONE*, vol. 11, p. e0167331, 12 2016.

[11] P. Bede, P. M. Iyer, E. Finegan, T. Omer, and O. Hardiman, "Virtual brain biopsies in amyotrophic lateral sclerosis: Diagnostic classification based on in vivo pathological patterns," *NeuroImage: Clinical*, vol. 15, pp. 653–658, 2017.

[12] T. Li, J. Howells, C. Lin, N. Garg, M. Kiernan, and S. Park, "8. predicting motor disorders from nerve excitability studies," *Clinical Neurophysiology*, vol. 129, p. e4, 4 2018.

[13] A. Sarica, A. Cerasa, P. Valentino, J. Yeatman, M. Trotta, S. Barone, A. Granata, R. Nisticò, P. Perrotta, F. Pucci, and A. Quattrone, "The corticospinal tract profile in amyotrophic lateral sclerosis," *Human Brain Mapping*, vol. 38, pp. 727–739, 2 2017.

[14] Y. Qi, C. Yang, H. Zhao, Z. Deng, J. Xu, W. Liang, Z. Sun, and J. D. V. Nieland, "Neuroprotective effect of sonic hedgehog mediated pi3k/akt pathway in amyotrophic lateral sclerosis model mice," *Molecular Neurobiology*, vol. 59, pp. 6971–6982, 11 2022.

[15] X. Luo, X. Kang, and A. Schönhuth, "Predicting the prevalence of complex genetic diseases from individual genotype profiles using capsule networks," *Nature Machine Intelligence*, vol. 5, pp. 114–125, 2 2023.

[16] N. Daneshafrooz, M. B. Cham, M. Majidi, and B. Panahi, "Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the wgcna and lasso algorithms," *Scientific Reports*, vol. 12, p. 20144, 11 2022.

[17] W. R. Swindell, C. P. S. Kruse, E. O. List, D. E. Berryman, and J. J. Kopchick, "Als blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia," *Journal of Translational Medicine*, vol. 17, p. 170, 12 2019.

[18] W. van Rheenen, F. P. Diekstra, O. Harschnitz, H.-J. Westeneng, K. R. van Eijk, C. G. J. Saris, E. J. N. Groen, M. A. van Es, H. M. Blauw, P. W. J. van Vught, J. H. Veldink, and L. H. van den Berg, "Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study," *PLOS ONE*, vol. 13, p. e0198874, 6 2018.

[19] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, January 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/