# Exploiting Machine Learning And Gene Expression Analysis in Amyotrophic Lateral Sclerosis Diagnosis

Hai-Long Nguyen
*Data and Intelligent Systems Laboratory*
*Posts and Telecommunications*
*Institute of Technology*
Hanoi, Vietnam
longnh.b23kd038@stu.ptit.edu.vn

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

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

***Abstract*—Despite many research efforts, the biological insight related to Amyotrophic Lateral Sclerosis (ALS), a rare disease resulting in the loss of motor neurons and causing mortality, remains elusive and leads to challenges to the diagnosis of the disease. Fortunately, gene expression data has recently appeared as a potential approach for the functionality analysis of genes related to orphan diseases, and for providing more accurate diagnosis outcomes. Moreover, with the explosion of machine learning (ML), implementing ML in analyzing biomedical data has become a promising direction with a noble effect on our lives. Leveraging these advantages, in this paper, we investigate to shed light on the effects of gene markers on ALS diagnosis and propose a novel gene combination that is effective in ALS diagnosis. We retrieve the datasets and perform the cleaning and pre-processing methods to obtain robust data for analysis. Then, the Max-Min Parents and Children (MMPC) and Sequential Forward Feature Selection (SFFS) algorithms are applied to achieve the optimal gene subsets effective for the final intelligent diagnosis model. Notably, the coefficient of the Ridge Classifier is utilized as the crucial score for determining the gene importance ranking table based on the selected gene signatures. All the possible gene combinations are evaluated and optimized in a set of robust machine learning algorithms. Consequently, a set of 20 genes identified through the Support Vector Machine (SVM) algorithm is selected as the optimal for the ALS diagnosis with an accuracy of 88.30% and an AUC score of 91.11%, which is dominant in comparison with notable traditional methods under the same datasets.**

***Index Terms*—Machine Learning, Gene expression, Gene Selection, Sequential Forward Feature Selection, Amyotrophic Lateral Sclerosis**

## I. Introduction

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a fatal and incurable neurodegenerative disease characterized by the progressive degeneration of motor neurons in the human nervous system, resulting in symptoms such as stiff muscles, muscle twitching, and gradual muscle atrophy, ultimately leading to death within 1-5 years [1], [2]. Despite the rarity of ALS, i.e., the prevalence of ALS in Europe is relatively low, with a ratio of over 3 cases per 100,000 people and even it's significantly lower in Asia, at approximately 0.7-0.8 cases per 100,000 individuals, global ALS cases are projected to reach nearly 370,000 by 2040 [3], [4]. However, the disease's etiology remains mostly unknown, posing challenges for effective treatment and early detection, given its genetic complexity and diverse pathology, necessitating tailored diagnostic and pharmaceutical approaches [5].

Since ALS still does not have a specific treatment method, early detection and managing its symptoms can substantially enhance the survival time of a patient. In [6], NAD+ is revealed to significantly improve the clinical features of ALS patients and lead to a potential treatment method for ALS [7]. These elucidate the life-saving potential of effective techniques and resources for forecasting the prevalence and incidence of ALS.

Meanwhile, the utilization of machine learning techniques is regarded as one of the most efficacious methods for analyzing the biological functions of medical data, significantly enhancing diagnostic efficiency. A large number of studies have been done with the purpose of identifying effective biomarkers for the treatment and diagnosis of ALS. However, most current works only depend on the clinical data, which significantly varies between ALS patients; hence, it is hard to give reliable results. In [8], the author uses the biomedical image related to the ALS disease with appropriate signal processing techniques to identify the image related to ALS with acceptable results. Additionally, in [9], the authors introduced a sophisticated approach utilizing integrated image metrics to create a diagnostic model. This model's performance is solely dependent on clinical features [10], [11], achieving an average accuracy score of 90%. Nevertheless, the latest model encounters challenges due to its restricted dataset size, resulting in a lower true positive rate and a significant rise in false negative rates. This phenomenon may be attributed to the imbalance within the ALS dataset, given its classification as a rare disease. Moreover, further exploration is essential to gather additional insights for refining a resilient diagnostic model and enhancing the management of ALS treatment.

As mentioned, the ALS disease is genetically related, and approximately 5-10% of cases have a family history of this disease. The potential of applying genome data to research the biomarkers of ALS is high. In [12], [13], there are 30 genes revealed as contributing significantly to ALS, and some mutation genes are also related to ALS. Although there is substantial knowledge about ALS, its definitive causes and underlying mechanisms remain elusive. Much of the ongoing research in genomic analysis aims to pinpoint the most reliable biomarkers associated with ALS and elucidate its pathogenic mechanisms. However, despite advancements in genomic technology and the availability of diverse gene datasets, current research on ALS has encountered shortcomings in this regard. Leveraging the efficiency of the explosion and efficiency of genomic data and ML and DL, in [14], using the individual genome profiles, the authors proposed a novel method using a Capsule Network combined with PCA decomposition for developing an ALS diagnosis model. Moreover, in [15], a combination of WGCNA and LASSO regression algorithm is applied to ALS gene expression data to reveal a subset of 5 genes that is effective in the diagnosis of ALS. Additionally, 850 genes and 468 principal components were proposed in [16] as effective methods for ALS diagnosis using the SVM algorithm. Recently, [17] proposed an effective gene markers subset with ML models to diagnose ALS with 22 gene markers and a Logistic Regression algorithm through a novel gene selection and optimization procedure. In [18], the author proposed a novel method using the Statistically Equivalent Signature (SES) gene selection with XGBoost and Random Forest algorithm to reveal a set of 473 genes for ALS classification. Moreover, using several machine learning models in [19] in gene expression data does not produce sufficient results in classifying between ALS and ALS mimic samples. Likewise, the computational cost and complexity of the previous studies are also big concerns. Consequently, a critical research question is how we can develop an efficient procedure method for identifying the potential markers for diagnosis of ALS as well as less effort in the computational resources.

In this work, to cope with that problem, we propose a novel gene selection procedure based on the statistical analysis approach and machine learning algorithm to provide an efficient model and feature subset for ALS diagnosis. The MMPC algorithm in combination with Ridge regression and Sequential Forward Feature Selection (SFFS) algorithm is utilized to figure out the most effective gene combination for diagnosing ALS. Additionally, seven typical machine learning algorithms are trained and optimized using gene expression values of the selected genes to determine the optimal model. Numerical experiments are performed to verify the effectiveness of our proposed approach compared to the notable traditional methods. The obtained results prove that our method significantly outperforms the comparable approaches.

## II. Datasets and Pre-processing

### A. ALS Gene Expression Datasets

In our research, two well-known datasets including GSE112676 and GSE112680 [19] are used as the main resource for our analysis and development of an ALS diagnosis model. These datasets were generated from whole-blood gene expression profiles of ALS patients with Illumina HumanHT-12 V3.0 expression bead chip (the GSE112676) and Illumina HumanHT-12 V4.0 expression bead chip (the GSE112680). The different platforms would produce various results in the number of representation genes, and the probe mapping for each dataset will be varied. Hence, the datasets are processed individually and combined for further processing to ensure the developed method can work properly with both platforms.

Moreover, the expression gene data of two datasets are extracted from the microarray data via the *GEOquery* package from the R library for gene expression analysis. The detailed information of the sample, such as *Sex. SpO2, Annotation,.etc* gathered from the *series matrix* file corresponding to each dataset. Because our main focus is to analyze the effect of gene markers on ALS patients and to develop the optimal ALS diagnosis model, only the gene expression data and the annotation of each sample are considered.

### B. Probe Filtering And Pre-processing

Probes corresponding to gene symbols in both datasets are analyzed and the particular probes with the following criteria are eliminated: (1) There are no specific gene symbols or meaningless symbols assigned to this probe; (2) The probes do not correspond to the gene symbol. Based on the annotation provided for each platform, it is used for probe mapping to gene symbols. It is noteworthy to notice that the probes associated with multiple gene symbols are discarded. Meanwhile, the mean expression value of the gene symbol assigned to multiple probes will be considered for further analysis. Due to the difference in the dataset retrieval platforms, only the gene markers included in both datasets are examined. The expression data in two datasets with annotations are then combined for the next analysis step.

In addition, it is critical to mention that, under this research target, only the *Control* and *ALS* samples are investigated and analyzed while the *ALS-mimic* samples are neglected. The gene expression values are normalized using the *Standard Scaler* algorithm to get the mean of the normalized data at 0 and the standard deviation of 1. The obtained dataset comprises 1,042 samples, with 397 individuals diagnosed with ALS and 645 categorized as normal. The processed data is then divided into *Training* and *Testing* subsets with the appropriate data proportion of 90% and 10% respectively. Note that the sample ratio of the *ALS* and *Control* in the two subsets are kept the same.

## III. Methodology

Fig. 1 shows our proposed method including three main steps: (1) Data preprocessing, (2) Gene selection and ML model, and (3) Model estimation. Firstly, data pre-processing is implemented to gather the gene expression data, remove the outliers, filter the probes, and process the expression data to become more robust for the gene expression analysis. Additionally, the MMPC feature selection is also implemented in this phase to select the most effective gene markers for the

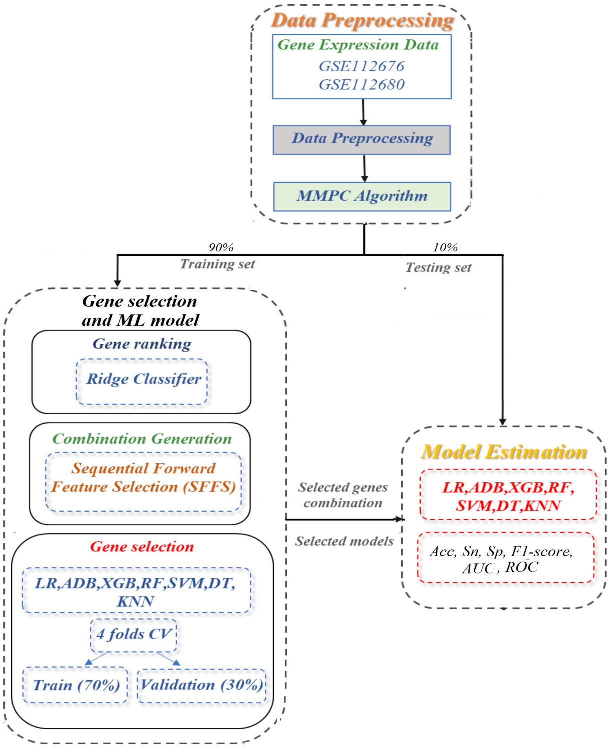

Fig. 1. Flow method

diagnosis of ALS. Then, the robust dataset with appropriate importance genes is fed into the second phase *Gene selection and Model*. In this step, the selected gene signature was re-ranked by the coefficient of the Ridge Classifier before applying the Sequential Forward Feature Selection (SFFS) to construct the possible gene combination. The optimal gene subsets were selected based on the result of the 4-fold cross-validation procedure on the training set. Finally, the selected gene combination corresponding with each machine learning algorithm is estimated for the performance on the testing set in the *Model Estimation* phase. The gene combination corresponding to the ML algorithm that provided the highest results in terms of accuracy and AUC score is considered the optimal gene subset and effective algorithm for ALS diagnosis.

On the other hand, seven typical machine learning algorithms, which are Logistic Regression (LR), Adaboost (ADB), XGBoost (XGB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbor (KNN), are investigated to estimate the effectiveness of the chosen gene combination for developing the most proper ALS diagnostic model.

### *A. Gene Selection Method*

The processed dataset includes 21029 genes, a significant number of genes. Such a large dataset may contain irrelevant genes that can degrade the performance of the ALS diagnosis model. From a mathematical point of view, the huge number of genes of individual patients leads to a higher correlation between genes, and more redundancy information will be added to the intelligent diagnosis model. Moreover, the high correlation and redundancy of information from the gene expression data can cause *overfiting* - a critical problem in training machine learning models, and directly reduce the effectiveness of the predictive model. Hence, it is crucial to reduce the number of irrelevant genes, which helps to improve the diagnosis performance and enhance the inference effectiveness in terms of algorithm complexity in training and optimizing hyperparameters.

Furthermore, leveraging the advantages of the MMPC algorithm, a constraint-based feature selection algorithm that works efficiently with datasets having huge feature spaces, like biomedical datasets, [20], is implemented to identify the most effective features for the diagnosis. The core MMPC mechanisms are performing multiple conditional independence tests and progressively excluding irrelevant and redundant variables. The final variables that remain from the whole procedure are considered as the MMPC output signature; in our problem, it is the selective gene signatures. To implement the MMPC algorithm, the MXM library provided by the R package is used [21].

### *B. Machine Learning-based Re-ranking Gene Score*

In this stage, the selected genes from the MMPC algorithm are investigated to re-rank based on the absolute coefficient of each gene in the classifier model. Our main purpose is to determine the most important gene markers that performed best for classification; hence, the Ridge classifier, a variant of the Ridge Regression, is implemented. This algorithm first converts the binary labels to $\{-1, 1\}$ and then treats the problem as the regression task. It is noteworthy to mention that the *L2*-regularization used by the Ridge Classifier to reduce the *overfitting* by adding a term controlled by $\alpha$ - the regularization strength. The *Mean Square Error* is also applied in this algorithm. By estimation and from the heuristic point of view and data distribution analysis, the Ridge classifier is considered an optimal choice for estimating the effect of each feature-gene marker into the final prediction of the machine learning model. It is important to note that hyperparameter tuning is also implemented in this phase to achieve the optimized model. The absolute coefficients of the final classifier are utilized as the gene importance score.

### *C. Marker Genes Discovery and Estimation*

After two feature selection phases, the remaining genes are considered the most informative gene markers to form the effective ALS diagnosis model. Based on the absolute coefficient scores obtained by the Ridge classifier, these genes are ranked in descending order to create a gene ranking table. SFFS algorithm is deployed afterward to determine gene combination sets. All the generated gene combinations are then fed into all seven considered machine learning algorithms to estimate their performance for determine the optimal gene subset. Considering all $k$ combinations, a total of $7 \times k$ models are taken into account. The hyperparameter tuning with each classifier corresponds to the gene subset to overcome the overfitting problem. The grid search procedure comprised of the 4-fold cross-validation is implemented to select the optimized configuration. It is notable clearly that in our implementation process of cross-validation, the training and testing are divided by a 70-30 ratio, and the proportion of the ALS and normal samples also follows this percentage. The optimal gene subset corresponding with the classifier that produces the best accuracy result is then evaluated on the separated testing set.

### *D. Diagnosis Model Development*

The selected gene subset corresponding to each ML algorithm is then estimated for the diagnosis performance on the testing set. Different ML algorithms are trained using the entire training dataset with the optimal gene combination. The hyperparameter tuning is also applied in this phase. The most efficient algorithm for the ALS diagnosis with the optimal gene subset is chosen based on the superior performance exhibited by the proper ML models.

## IV. Experimental results and discussion

Major evaluation metrics are utilized to estimate the performance of the ML model in our numerical experiments, including Accuracy (Acc), F1-score (F1), and AUC-ROC score (AUC). The AUC and Acc scores measure the precision of the intelligent model while the F1 score estimates the ability of the ML algorithm to balance the classification of two labels.

### *A. Gene Expression Data Processing*

The gene expression data is extracted from two datasets, GSE112676 and GSE112680, consisting of 1042 samples with 397 ALS and 645 normal and 21029 two-dataset-overlapped genes are used for gene expression analysis. After performing the probe mapping process and the data cleaning method to remove the outliers and redundant data, the number of remaining genes from GSE112676 and GSE112680 are 21571 genes and 21621 genes, respectively. The average expression value replaces the probes having the same corresponding gene symbols. The combination of the two datasets is then normalized and fed into the gene selection procedure with the MMPC algorithm. To get the optimal result, we implement the MMPC with the $max\_k$ values of the conditioning sets to use is 3, and the threshold for the statistic level of significance is $0.1$. Besides, the statistically independent testing algorithm used in the MMPC is set for using *testIndFisher*. Finally, only 24 genes from a total of 21029 genes are chosen as the most important features. In other words, the selected 24 genes are highly important in relation to the ALS and control samples.

### *B. Gene Importance Score Analysis*

The combination of 24 genes found by the MMPC feature selection is fed into the Ridge Regression model to evaluate the coefficient value of each gene that affects the performance of the ALS diagnosis. The hyper-parameter tuning is crucial with the regularization strength - $\alpha$ and the *solver* methods. The Ridge Classifier is employed by using the *scikit-learn* model and applying the *Grid Search* algorithm with a cross-validation to obtain the optimized model. Obviously, the indexes of the gene symbols provided by the MMPC feature selection algorithm are interchanged to adapt to the optimal machine learning model rather than the normal statistic models as the MMPC. Additionally, one of the characteristics of the Ridge Regression model is that it does not reduce the coefficient of the features in the model to zero; hence, it preserves all the data information for the final decision. Based on the coefficient absolute value, the ranking table of gene coefficient scores is established.

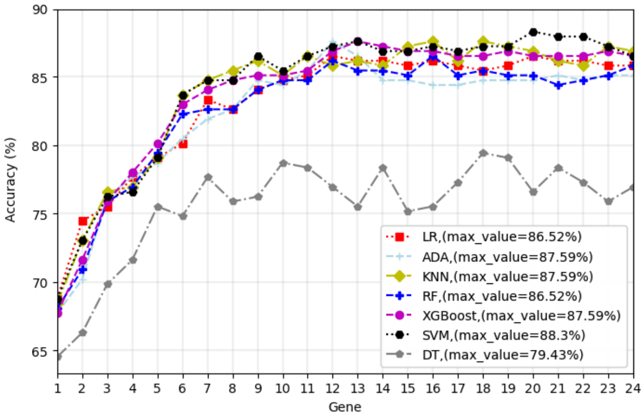

Fig. 2. Gene combination estimation

### *C. The Optimal Gene Subset Selection*

Using the gene ranking table from the gene importance score analysis process, we investigated the effect of each gene combination on the ML models. A total of 24 gene subsets are constructed from the selected gene markers to figure out the most effective gene subset. Seven typical machine learning algorithms and 24 gene combinations are investigated. As a result, 168 models are constructed to search for the best gene combination corresponding to the ML models. The 4-fold cross-validation procedure is implemented to tune the hyper-parameter and to estimate the efficiency of the diagnosis models. Fig. 2 depicts the obtained results of the seven ML models, which produce the highest accuracy score corresponding to each associate with the optimal gene combination. The results show that the SVM achieves the highest accuracy, 88.3% with the optimal subset of 20 genes. Besides, KNN, Adaboost, and XGboost offer the same accuracy at 87.59% while the numbers of genes are 16, 12, and 13, respectively. The details of the validation results are summarized in Table. I.

### *D. Diagnosis Performance Validation*

To estimate the effectiveness of each ML algorithm corresponding with the optimal subset, the diagnosis performance of each model is evaluated on the testing set. Seven considered ML models are trained on the entire training set, and the final performances on the testing set are evaluated. Note that hyper-parameter tuning is also implemented with the *Grid Search* to

TABLE I
VALIDATION RESULTS OF THE OPTIMAL GENE COMBINATIONS

| ML algorithm | Parameters | | | | | | |
|---|---|---|---|---|---|---|---|
| | $N$_genes | AUC (%) | Precision (%) | Sn (%) | Sp (%) | F1-score (%) | Acc(%) |
| *LR* | 12 | 91.44 | 84.15 | 79.43 | 90.85 | 81.73 | 86.52 |
| *Adaboost* | 12 | 91.23 | 82.14 | 85.98 | 88.57 | 84.02 | 87.59 |
| *KNN* | 16 | 89.01 | 86.00 | 80.37 | 92.00 | 83.09 | 87.59 |
| *RF* | 16 | 90.66 | 81.65 | 83.18 | 88.57 | 82.41 | 86.52 |
| *XGB* | 13 | 91.83 | 83.33 | 84.71 | 89.71 | 83.72 | 87.59 |
| *SVM* | 20 | 91.11 | 84.26 | 85.05 | 90.29 | 84.65 | 88.3 |
| *DT* | 18 | 78.34 | 72.48 | 73.83 | 82.86 | 73.15 | 79.43 |

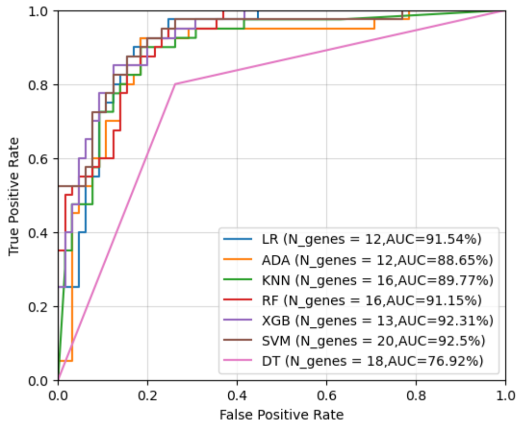

Fig. 3. ROC-AUC score on the testing set

determine the optimized model along with its optimal gene combination. Fig. 3 illustrates the AUC scores of the ML algorithms with their appropriate optimal gene subset on the testing set. The XGB and SVM algorithms achieve the highest AUC scores of 92.31% and 92.50%, followed by the LR, RF, KNN, and Adaboost with that of 91.54%, 91.15%, 89.77%, and 88.65%, respectively. However, the DT algorithm provides the worst performance with the AUC score of 76.82% only. In addition, as shown in Fig. 4, XGB, LR, and SVM obtain the highest accuracy on the testing set, with a score of 84.76%. Considering the highest accuracy and AUC scores on the testing set, we propose a combination of 20 genes with the SVM algorithm as the most effective gene markers for the ALS diagnosis. To verify the performance of our proposed method, it is compared to three notable traditional works using the same dataset [15]–[17], and the most state-of-the-art results are summarized in Table. II.

TABLE II
PERFORMANCE COMPARISON

| Comparable methods | Performance parameters | | | | |
|---|---|---|---|---|---|
| | $N$_genes | AUC (%) | Sn (%) | Sp (%) | Acc (%) |
| [15] | - | 86.50 | - | - | - |
| [16] | 850 | - | 86.00 | 87.00 | 87.00 |
| [17] | 22 | 87.90 | 72.86 | 88.94 | 82.28 |
| *This work* | **20** | **91.11** | **85.05** | **90.29** | **88.30** |

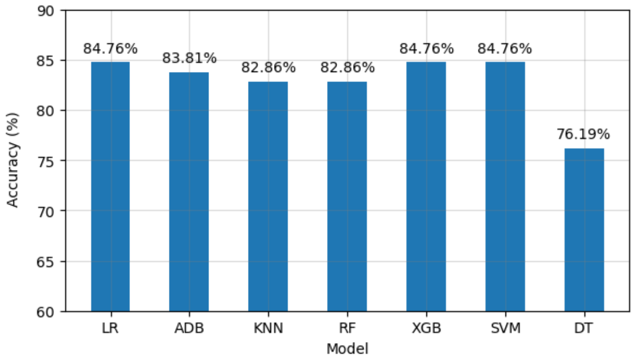

Fig. 4. Accuracy on testing data

### *E. Discussion*

We developed a robust method to select the efficient gene markers for the ALS diagnosis based on a three-phase ML-based gene selection procedure. The optimal 20 genes subset and the SVM algorithm are determined as the optimal configuration for the ALS diagnosis with the cross-validation results of 88.30% accuracy and 91.11% AUC score. Our proposed gene subset is significantly small compared to other existing works, as [16] proposed with 850 genes and [17] with 22 genes. Moreover, our proposed model also outperforms the model given in [15] in terms of the AUC score. The reason behind the success of our method comes from the efficient feature selection process utility of three procedures of gene number reduction that summarize the pros and cons of three types of feature selection algorithms, including filter, wrapper, and embedding methods. In our approach, the number of genes drops drastically thanks to applying the MMPC feature selection algorithm. The statistical testing implemented on the MMPC algorithm estimates the effect of each gene in the set of 21029 genes on the ALS diagnosis. Moreover, the MMPC is also known as an efficient method for processing high-dimension data. The subset of 24 genes, which is the result of the MMPC feature selection procedure, is then fed into the Ridge Classifier to estimate the effect of each feature on the decision of the ML algorithm. Based on the characteristics of the Ridge Regression that are different from the Logistic Regression or Linear regression, they do not eliminate the feature information by preventing assigning the coefficient score to zero. Hence, all gene markers are still considered after the filtering method rather than the approach with LASSO regression in [17]. The lower coefficient score indicates the less importance of this gene in the final decision. The coefficient is then re-ranked from highest to lowest to form the ranking table for the sequential forward feature selection process. On the other hand, the SFFS algorithm is applied intensively to search the optimal gene subset for developing a final diagnosis model of ALS disease. Seven population ML algorithms combined with the cross-validation procedure were used to search for the optimal configuration of each algorithm with appropriate gene combinations. It is important to mention that the way we set up the cross-validation procedure also contributes significantly to our final results.

## V. Conclusion

In this work, we investigate the effectiveness of the machine learning algorithms with the optimal gene subsets for the ALS diagnosis. To achieve the optimal gene combination for the diagnosis of ALS, we propose a novel and effective sequential method for feature selection that combines the MMPC algorithm, the Ridge classifier, and the Sequential Forward Feature Selection procedure. There were 24 gene combinations produced from the feature selection process, which then estimated performance with seven machine learning algorithms. As a result, a combination of 20 genes according to the SVM algorithm produces the most accurate result in a diagnosis of ALS with an AUC score of 91.11% and an accuracy of 88.30%. Compared with other work in the same dataset, our proposed diagnosis model outperforms the novel existing studies in terms of accuracy and the number of genes required to perform the diagnosis model. Our study introduced an efficient approach to selecting genes and developing an accurate predictive model for ALS disease. This approach has the potential to serve as a foundation for both clinical diagnostic evaluations and comprehensive investigations into biological mechanisms.

## References

[1] M. K. Kotni, M. Zhao, and D.-Q. Wei, "Gene expression profiles and protein-protein interaction networks in amyotrophic lateral sclerosis patients with C9orf72 mutation," *Orphanet Journal of Rare Diseases*, vol. 11, p. 148, 12 2016.

[2] A. Catanese, S. Rajkumar, D. Sommer, P. Masrori, N. Hersmus, P. Van Damme, S. Witzel, A. Ludolph, R. Ho, T. M. Boeckers, and M. Mulaw, "Multiomics and machine-learning identify novel transcriptional and mutational signatures in amyotrophic lateral sclerosis," *Brain*, vol. 146, pp. 3770–3782, 9 2023.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] J. E. de la Rubia, E. Drehmer, J. L. Platero, M. Benlloch, J. Caplliure-Llopis, C. Villaron-Casales, N. de Bernardo, J. AlarcÓn, C. Fuente, S. Carrera, D. Sancho, P. GarcÍa-Pardo, R. Pascual, M. JuÁrez, M. Cuerda-Ballester, A. Forner, S. Sancho-Castillo, C. Barrios, E. Obrador, P. Marchio, R. Salvador, H. E. Holmes, R. W. Dellinger, L. Guarente, and J. M. Estrela, "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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] N. Daneshafrooz, M. Bagherzadeh 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.

[16] 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.

[17] D.-L. Vu and H.-C. Le, "Machine Learning-Based ALS Diagnosis Using Gene Expression Data," in *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 354–359, IEEE, 12 2023.

[18] K. Founta, D. Dafou, E. Kanata, T. Sklaviadis, T. P. Zanos, A. Gounaris, and K. Xanthopoulos, "Gene targeting in amyotrophic lateral sclerosis using causality-based feature selection and machine learning," *Molecular Medicine*, vol. 29, p. 12, 1 2023.

[19] 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.

[20] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, pp. 31–78, 10 2006.

[21] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, and I. Tsamardinos, "Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets," *Journal of Statistical Software*, vol. 80, no. 7, 2017.