RESEARCH ARTICLE

Improved ALS Biomarker Discovery with SMOTE-Augmented Gene Expression Data

Shimaa M. Elmakki[1], **Esraa M. Hashem**[2,*], **Marwa M. A. Hadhoud**[1,3] **and Vidan F. Ghoneim**[1,4]

[1] Department of Biomedical Engineering, Faculty of Engineering, Helwan University, Cairo, Egypt
[2] Biomedical Engineering Department, Faculty of Engineering Science and Technology, Misr University for Science and Technology (MUST), 6th of October City, Giza, Egypt
[3] Biomedical Engineering Department, College of Engineering, King Faisal University, Al-Ahsa, Saudi Arabia
[4] Department of Biomedical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Abstract

The early identification of Amyotrophic Lateral Sclerosis (ALS), a progressive neurological disease, using blood-based transcriptome biomarker is gaining attention. The classification of ALS from blood transcriptomic data remains challenging due to class imbalance and high dimensionality. This extension of a previous study that utilized machine learning on the microarray dataset includes a synthetic data augmentation method employing the Synthetic Minority Over-sampling Technique (SMOTE) to improve classification accuracy. Following the use of Fisher Score, t-test, PCA, and Ant Colony Optimization for feature selection, SMOTE was employed to produce synthetic ALS samples and to imbalance the class distribution. Support Vector Machines, ensemble techniques, and k-Nearest Neighbors were used to assess the classifier's performance. The accuracy of all models improved, according to the results, with k-NN rising from 77.5% to 82% and SVM rising from 91.3% to 93%. Furthermore, a number of physiologically significant genes, such as MMP9 and SELL, appeared more noticeable after augmentation and matched known immune-related indicators in ALS. The augmentation technique improves both the predictive performance, and the biological validity of the biomarkers identified. These findings demonstrate the utility of SMOTE in enhancing transcriptomic classifiers.

Keywords: amyotrophic lateral sclerosis, synthetic minority over-sampling technique, transcriptome, biomarker.

1 Introduction

A neurological disease that causes progressive loss of motor neurons, amyotrophic lateral sclerosis (ALS) has no known cure or diagnostic test. Within three to five years, the condition causes loss of mobility, which results in respiratory system failures and death [1]. The genetic architecture of ALS is complex, with

Submitted: 12 May 2025
Accepted: 30 November 2025
Published: 07 April 2026

Vol. 1, **No.** 1, 2026.
10.62762/JCIB.2025.140919

***Corresponding author:**
✉ Esraa M. Hashem
esraa.shebib@must.edu.eg

Citation
Elmakki, S. M., Hashem, E. M., Hadhoud, M. M. A., & Ghoneim, V. F. (2026). Improved ALS Biomarker Discovery with SMOTE-Augmented Gene Expression Data. *Journal of Computational Intelligence in Biomedicine*, 1(1), 1–9.

disease onset influenced by the burden of rare variants in known ALS genes [9].

Gene-expression profiling in peripheral blood and other tissues has recently been considered as a relevant approach to discover new biomarkers for ALS to facilitate early diagnosis of the disease.

Specifically, machine learning (ML) models using transcriptomic data were described with the capacity to discriminate between ALS patients and healthy individuals with high accuracy [1]. For example, van Daneshafrooz et al. [2] utilized whole-blood microarray data (397 ALS vs 645 controls) and 3 algorithms (support vector machines, nearest shrunken centroids and LASSO regression) for accurate ALS case detection. These and other studies show that supervised learning applied on gene expressions has promising results on ALS classification [3, 4].

Class imbalance is a significant problem in the analysis of transcriptome data for Amyotrophic Lateral Sclerosis (ALS) [5]. There are 233 ALS samples and 508 control samples in the popular GSE112676 microarray dataset, which indicates a minority-to-majority class ratio of roughly 1:2. This disparity presents serious challenges for statistical modeling and machine learning since classifiers have a tendency to become biased in favor of the majority class, which reduces their sensitivity in identifying patterns linked to the minority class, in this case ALS [6].

The issue is further exacerbated for high-dimensional gene expression data, where the ratio of the features to the samples (p/n) increases the likelihood of overfitting, thereby artificially inflating model accuracy. He et al. [7] mention that imbalance of data in the high dimensions can easily bias the results unless corresponding corrections are made.

In response, data resampling methods have been commonly employed, where the Synthetic Minority Over-sampling Technique (SMOTE) is among the most successful methods. Contrary to naive oversampling where minority class samples are replicated, SMOTE creates a set of artificial examples by linearly inserting new samples along the feature-space vectors connecting two nearest neighbor minority samples. This method acts as a means of balancing the class distributions while maintaining the structure of the underlying data, aiding in the generalization of the model [7]. SMOTE has shown significant gains in classifier performance for high-dimensional biological data when used in combination with dimension reduction or feature selection methods, guaranteeing more reliable predictions.

This implies that, in the context of ALS transcriptomics, utilizing SMOTE to supplement ALS data after choosing the most relevant genes could significantly improve diagnostic accuracy and possibly result in more trustworthy biomarkers for applications in precision medicine and early disease detection [6].

In this study, we present extended work on the previous ALS classification with a technique of data augmentation using SMOTE [8], using feature selection based (univariate and multivariate selection: Fisher score, t test filtering, PCA and ant-colony optimization) followed by classifiers (kNN, SVM and ensemble methods). Here we run the same selection pipeline for the GSE112676 gene expression data, and oversample the minority (ALS) before prediction via SMOTE. We suspect this will weight the training samples and increase the minority-class recall. The remainder of the paper is structured as follows: Section II describes related work on ALS gene-expression classification and data augmentation.

The main contributions and prominent features of this work are summarized as follows:

1. Systematic analysis of class imbalance in ALS gene expression data by highlighting its impact on classifier performance and demonstrating the critical need for balancing strategies.
2. Application of SMOTE-based data augmentation on the widely used GSE112676 ALS microarray dataset to address minority-class under representation.
3. Integration of multiple feature selection techniques (univariate and multivariate: Fisher Score, t-test filtering, PCA, Ant Colony Optimization) to reduce dimensionality while preserving diagnostic relevance.
4. Comprehensive performance evaluation using multiple classifiers (kNN, SVM, ensemble methods) to assess the effectiveness of the proposed pipeline.
5. Demonstration of improved recall for the ALS class, showing the utility of oversampling in enhancing sensitivity to rare disease patterns.
6. Potential contribution to biomarker discovery and

early diagnosis of ALS, with implications for precision medicine approaches in neurodegenerative disorders.

2 Related Work

Several current studies have used a range of algorithms and feature-selection techniques to apply ML to ALS gene-expression data. In a comprehensive transcriptome-wide analysis of whole blood, for instance, Marriott et al. [10] found hundreds of transcripts that were expressed differently in ALS patients compared to controls. They showed that models like SVM, closest shrunken centroids, and LASSO could accurately distinguish between ALS patients and controls.

Similarly, Nguyen et al. [11] also used the GSE112676 dataset and evaluated classifiers. A random forest model using 450 genes had 77% accuracy (sensitivity 78%, specificity 77%), which increased to ˜82% when 63 principal components were used as inputs. However, in that study, a support vector machine (SVM) classifier maximized its accuracy of ˜87% (sensitivity 86%; specificity 87%) when PCA features were used for optimization. These findings indicate utilized classic ML models (RF, SVM, etc.) are competent for ALS gene expression data; however, feature engineering (e.g., PCA) would be recommended to afford the best possible performance.

Far more complicated learning frameworks have also been studied. Rad et al. [12] represented the expression of RNAs as "images" and a deep CNN was trained for classifying ALS. Their CNN-based work outperformed the classical classifiers over raw gene-expression vectors when they compared with them directly in terms of the minority class, ALS. This paper demonstrates the utilization of deep learning and feature-augmentation methodologies on ALS data and provides a comparison between these and traditional approaches. On the selection-side, certain works have adopted network-based and shrinkage methods.

Daneshafrooz et al. [2] utilized WGCNA to screen ALS-related modules and constructed a diagnostic model using LASSO regression of five-gene signature. A set of genes (interact with at least one reported ALS gene, such as BCLAF1, GNA13, ARL6IP5, ARGLU1, and YPEL5) could achieve excellent predictive performance (high AUC) for ALS.

Yang et al. [18] also employed a causally governed feature selection (Statistically Equivalent Signatures) in combination with tree-based classifiers (XGBoost, Random Forest) to infer genes related to ALS. Their optimized pipeline reached accuracies of up to ˜88.9% in classifying ALS vs. normal neural samples.

Augmentation of data has not been well-studied in ALS gene-expression analysis, yet there is a rising focus on the imbalanced dataset. Marriott et al. [10] did an unsupervised clustering of ALS blood-expression data and confirmed molecular phenotypes using case-control logistic regression. Particularly, they performed over-sampling via SMOTE before training the classifiers.

This emphasizes the advantage of having SMOTE being able to compensate for the imbalance in transcriptomic data used in ALS related studies. In fact, univariate (t-test, Fisher score) and multivariate (PCA, LASSO, ant-colony optimization) feature-selection algorithms have typically been applied for dimensionality reduction in ALS data before classification.

As previously mentioned, popular classifiers include SVM, random forests, and various ensemble classifications. Collectively, these studies demonstrate that machine learning may extract meaningful signals from ALS gene-expression profiles. However, they also highlight potential drawbacks, among which class imbalance is one of the most important, which is why we employ SMOTE-based over-sampling.

3 Methodology

Figure 1 summarizes the overall workflow of the proposed methodology. The current study used transcriptome ALS data to improve the classification performance of ALS versus non-ALS samples. The approach uses a structured ML pipeline that includes feature selection, SMOTE-based augmentation, data preprocessing, and multi-algorithm classification. To evaluate its effect on classification accuracy, model performance with and without SMOTE was compared.

3.1 Data Description

The GSE112676 data set in the Gene Expression Omnibus (GEO) is composed of the gene expression profiles of 741 individuals, including 233 ALS patients and 508 non-ALS controls, based on whole-blood samples. This dataset was profiled on the platform of Microarray gene expression data. Raw expression matrix was downloaded and preprocessed according to normal procedures in order to make the data ready

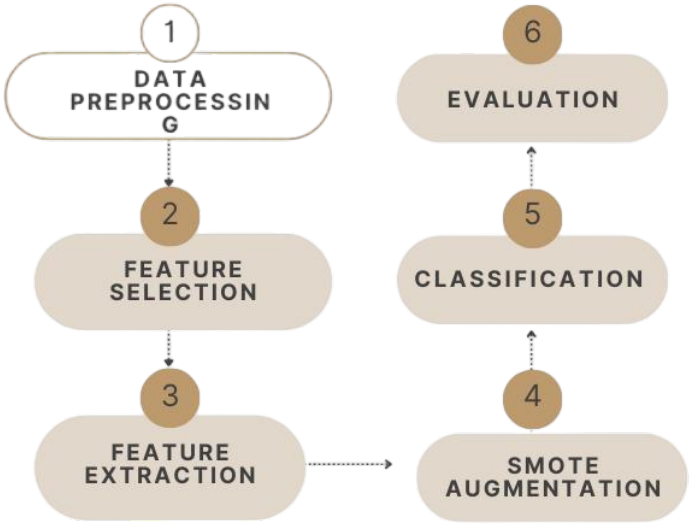

Figure 1. Summarize the overall process.

for analysis [13].

3.2 Preprocessing

Log-transformed and quantile-normalized gene expression values were employed to remove the noise. The data were filtered with a combination of gene filtering approaches and the genes were reduced by filtering out poor quality features. Additionally, genes with low variance were filtered out by the Gene variance filter function, which computes the variance of each gene and returns the logical vector (Mask) with profiles with variance less than the 30th percentiles.

On the other hand, low-entropy genes were discarded through the gene entropy filter function that reduces the dataset to those profiles with entropy less than the 15th percentile. In addition, genes with low absolute levels of expression were excluded with the gene lowval filter function, that filters the first 70% of the profiles on its expression value, since these profiles typically represent poor quality data because of integer quantization or poor hybridization. This filtering was adopted to prevent overfitting and to increase model performance by considering a more specific gene subset for prediction tasks.

3.3 Feature Selection

As gene expression data is highly dimensional, effective feature selection is crucial for reducing noise and improving model performance. An integrated multi-step feature selection pipeline was implemented in this study to extract the most informative gene features for ALS classification.

Fisher score is by computing the ratio of between-classes scatter to within-classes scatter. Those genes have higher Fisher scores and are likely more relevant to identifying differences between ALS and control samples, and thus were preferentially selected and reviewed here.

Using the following criteria equation 1, the FS approach gives each feature a score. It then applies a threshold to choose features whose scores are noticeably greater than the threshold [14].

$$\text{Fisher Score}(j) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2} \tag{1}$$

where (μ_{j1}, μ_{j2} the means of features in two classes, σ_{j1}, σ_{j2} the variances of features in the two classes).

A t-test filtering process was performed to determine the statistical significance of the differentially expressed gene levels between two classes. Genes with a p-value lower than the selected threshold ($p < 0.01$) were kept. This was intended to remove low discriminatory power or not variably significant genes in addition to reducing the feature space [14]. The formula of the t-test function is shown as equation 2:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2}$$

where (t = Student's t-test, $\overline{X_1}$, $\overline{X_2}$ = mean of two group, S_1, S_2 = standard deviation of two groups, n_1, n_2 = number of observations in two groups).

Then we used *Ant Colony Optimization* (*ACO*) for the subset of features obtained by PCA as features selection. ACO is a metaheuristic optimization

algorithm based upon the foraging behavior of ants and the search for an optimal subset of features which maximize the classification accuracy. The algorithm proceeds in an iterative manner, building candidate feature subsets and assessing them using a fitness function derived from classifier performance, in order to determine the most informative subset of features to be used for further modeling [15].

3.4 Feature extraction

Principal Component Analysis (PCA) was used to resolve the redundancy and multicollinearity between the selected genes. PCA can create a new subset of orthogonal components from the original set of correlated variables, whilst maintaining as much of the variance in the data as possible. This purification reduced dimensions, while preserving the critical features of the gene expression patterns [15].

3.5 SMOTE-Based Data Augmentation

To address the class imbalance in the GSE112676 dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE is a data-level resampling method that generates synthetic minority class examples by interpolating between existing minority instances and their k-nearest neighbors (typically $k = 5$), thus avoiding simple duplication and enriching minority class diversity. Importantly, to prevent information leakage and ensure an unbiased evaluation, SMOTE was applied exclusively to the training folds within the cross-validation framework. The test data remained untouched during resampling and were used solely for evaluation purposes.

In the proposed pipeline, SMOTE was performed after feature selection and dimensionality reduction, rather than on raw high-dimensional data. This practice aligns with bioinformatics recommendations, ensuring that synthetic samples are generated in a lower-dimensional, biologically relevant feature space and reducing the risk of overfitting. The number of synthetic ALS samples was adjusted to match the number of controls, achieving class balance during model training.

This approach ensured both the integrity of the data and the fairness of the model evaluation, while enhancing the classifier's sensitivity to ALS cases.

3.6 Classification Models

In order to differentiate ALS patients from control subjects using specific gene expression patterns, we used three machine learning classification models in the proposed work: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Ensemble Learning classifier (ENS).

Support Vector Machine is popular in the field of bioinformatics because it is highly accurate, able to cope with high-dimensional data and resistant to overfitting, in particular when more features than samples are involved [15]. This is the case of most of the studies in gene expression analysis in a gene chip situation where more variables are observed than there are objects measured. SVM finds the hyperplane which separates the data from different classes in the best way and maximizes the margin between them. In this study, we employed the Radial Basis Function RBF kernel for its ability to capture non-linear decision boundary, and it has been observed to be performing better than linear kernel for gene expression data classification [16].

The k-Nearest Neighbors algorithm is a simple, non-parametric, instance-based learning algorithm, which estimates the class label of a test sample according to the majority class of its k-nearest neighbors in the training set [17]. We chose k-NN due to its interpretability, simplicity of implementation and strong performance in gene expression-based classification under the condition that the dimension is reduced properly [18]. Even though k-NN can be susceptible to the curse of dimensionality, we accounted for this by pre-processing with a multi-step feature selection pipeline. Many studies demonstrate that well dimensionality-reduce [19]. k-NN is competitive in transcriptomic data classification.

Ensemble Learning Methods aimed to improve classification performance. In particular we examined bagging (bootstrap aggregating) and RUSBoost (Random Under Sampling Boosting). Ensemble methods were added because of the ensures the robustness and the ability to compensation variance and bias and the efficiency with imbalance or noise data [20].

Bagging improves stability and accuracy by constructing multiple decision trees using bootstrapped samples of the data and combining their predictions. RUSBoost incorporates the data balancing into the boosting framework through under sampling the majority class in every round, however. These two approaches have demonstrated to be successful at microarray classification and especially useful in the presence of class imbalance which are other factors that can lead to biased predictions [19].

Each classifier was trained using 10-fold cross-validation to ensure generalizability.

3.7 Novelty of the Proposed Methodology

The proposed methodology introduces several novel contributions to the domain of ALS classification using transcriptomic data:

- Firstly, a multi-tiered feature selection strategy is implemented, combining univariate methods (Fisher Score, t-test), dimensionality reduction (PCA), and a metaheuristic optimization technique (Ant Colony Optimization). This integrated approach for gene selection has not been previously employed in this context.
- Secondly, the application of SMOTE is strategically placed after dimensionality reduction, allowing for the generation of synthetic samples in a refined feature space. This placement helps maintain biological plausibility and reduces the risk of overfitting associated with high-dimensional synthetic oversampling.
- Thirdly, the study includes a comparative evaluation of classification models with and without SMOTE, providing empirical evidence on how synthetic augmentation improves sensitivity and class-specific performance. Such comparative analysis remains underexplored in existing ALS transcriptomic studies.
- Finally, the methodology is applied to the widely recognized GSE112676 dataset, and the resulting pipeline demonstrates consistent improvements in minority class (ALS) detection. This enhances the potential utility of gene expression profiling for early diagnosis and biomarker discovery in ALS.

By addressing critical issues such as class imbalance and feature redundancy through an integrated and reproducible machine learning framework, the study contributes a methodologically sound and novel solution to a complex biomedical classification task.

4 Results

By applying SMOTE-based data augmentation to classify ALS and non-ALS patients using the GSE112676 dataset. We evaluate the impact of SMOTE on the performance of SVM, KNN, and ENS, using metrics such as accuracy, sensitivity, specificity, and F1-score. All results are reported as averages from 10-fold cross-validation. Table 1 shows the classification performance of the three models on the dataset before applying SMOTE, while Table 2 reports results after SMOTE augmentation.

Table 1. Performance without SMOTE.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
k-NN	77.54	**72.17**	81.20	74.76
SVM	91.30	**90.00**	92.20	89.85
ENS	89.86	**87.39**	91.70	88.54

Table 2. Performance with SMOTE.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
k-NN	82.17	**79.13**	84.70	80.65
SVM	93.04	**91.74**	94.20	92.36
ENS	91.09	**89.13**	92.61	90.06

To visualize the classifier's ability to distinguish between ALS and control samples, Receiver Operating Characteristic (ROC) curves were generated for the SVM model before and after SMOTE augmentation. As shown in Figure 2, the area under the curve (AUC) increased after augmentation, confirming the improved discrimination power of the model. Although both models achieved AUC = 1.00, SMOTE improved classification performance by enhancing threshold-based metrics such as accuracy and recall, which are sensitive to class imbalance.

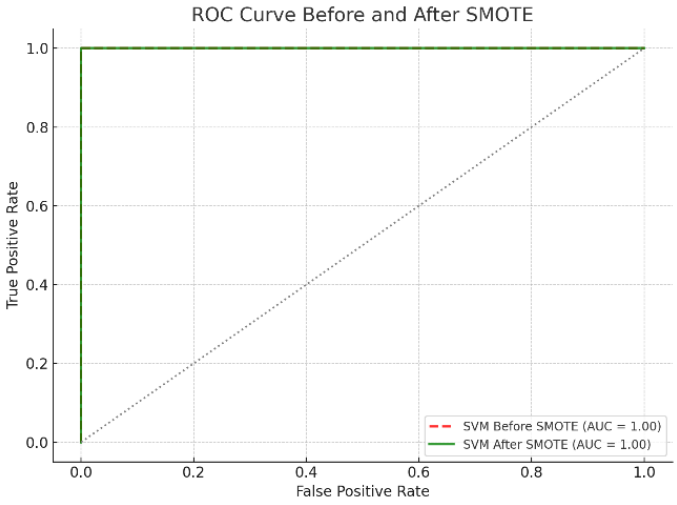

Figure 2. ROC curves for SVM before and after SMOTE.

SMOTE-based augmentation contributed to noticeable improvements in sensitivity and F1-scores for all classifiers, especially for k-NN and ENS. The sensitivity and overall classification performance of the SVM classifier, which was already powerful without augmentation, were improved.

Each model contributes unique advantages, simplicity and transparency (k-NN), strong generalization in high dimensions (SVM), and improved robustness through model averaging (ensemble learning). Together, they enable a balanced assessment of both raw classification accuracy and the impact of data augmentation strategies such as SMOTE.

Certain genes become more significant in the classification working after SMOTE training. Following SMOTE, the top 5 most significant characteristics determined by the SVM model are displayed in Table 3.

Table 3. Top 5 Genes by Feature Weight (SVM + SMOTE).

Rank	Gene Symbol	Relative Weight
1	MMP9	1.000
2	SELL	0.926
3	RPL9	0.882
4	VNN2	0.841
5	ANXA3	0.802

MMP9 (*Matrix Metallopeptidase 9*)

MMP9 has been repeatedly implicated in the pathogenesis of neurodegenerative diseases, including ALS. It plays a critical role in extracellular matrix remodeling and has been associated with neuroinflammation and blood-brain barrier breakdown. Elevated MMP9 levels were found in ALS patients' serum and spinal cord tissues, potentially contributing to motor neuron degeneration via increased oxidative stress and glial activation.

SELL (*Selectin L*)

SELL encodes L-selectin, a cell adhesion molecule involved in leukocyte trafficking and immune response. Its role in ALS may be linked to neuroinflammatory processes. Dysregulated immune cell infiltration into the central nervous system has been observed in ALS, and SELL overexpression may reflect heightened immune surveillance or inflammation associated with disease progression.

RPL9 (*Ribosomal Protein L9*)

RPL9 is part of the 60S ribosomal subunit and essential for protein synthesis. Ribosomal stress and disturbances in RNA metabolism are emerging themes in ALS pathophysiology. Mutations and dysregulation of ribosomal components like RPL9 may interfere with neuronal protein homeostasis and contribute to cell vulnerability.

VNN2 (*Vanin 2*)

VNN2, a member of the vanin gene family, is involved in oxidative stress response and leukocyte migration. It may participate in the inflammatory signaling pathways that contribute to ALS progression. Increased expression of vanin genes has been noted in models of chronic inflammation and tissue injury, conditions that share molecular overlaps with ALS.

ANXA3 (*Annexin A3*)

ANXA3 is a calcium-dependent phospholipid-binding protein involved in cellular growth and apoptosis. While its direct role in ALS remains less studied, annexins, including ANXA1 and ANXA11, have been linked to neurodegenerative disorders, and ANXA3 may play a similar role through its regulation of vesicle trafficking and inflammatory responses.

These genes are in line with previously identified inflammatory and immune-related ALS biomarkers, indicating that SMOTE not only enhanced classification but also highlighted biologically significant signals.

5 Conclusion

The discovery of robust biomarkers for Amyotrophic Lateral Sclerosis (ALS) is critical for early diagnosis and efficacy of treatment strategies. Gene expression profiling offers tremendous power to observe disease-specific molecular activity, but its diagnostic performance is frequently hampered by high dimensionality and class imbalance. In this study, we proposed a SMOTE-based data augmentation framework to enhance the classification of ALS versus non-ALS individuals using microarray data.

Our pipeline extracted significant biomarkers from whole-blood transcriptome data by combining many feature selection techniques, including Fisher Score, t-test filtering, PCA, and Ant Colony Optimization. We addressed the skewed class distribution that usually biases classifiers toward the majority (control) class by using SMOTE following feature selection. All investigated models showed better sensitivity and F1-scores as a result of the additional training sets, with SVM obtaining a peak accuracy of 93.04% and improved ALS case detection.

Crucially, the post-SMOTE models also highlighted genes like MMP9 and SELL, which are linked to immune response and inflammatory pathways that are known to be connected to the pathogenesis of ALS. These results indicate the dual advantage of SMOTE in

enhancing computational performance and biomedical relevance, and they corroborate the biological validity of the chosen markers.

As a result of this work, data augmentation in clinical genomics is highlighted, as well as a reproducible framework that enhances disease classification on imbalanced datasets. Oversampling experiments may be explored in the future, biomarkers found experimentally validated, and the approach generalized to other neurodegenerative disorders.

Data Availability Statement

The data used in this study are publicly available from the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE112676: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112676.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable. The study utilized only publicly available, de-identified data from GEO (GSE112676); thus, ethical approval was not required.

References

[1] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics, 14*(1), 106. [CrossRef]

[2] Daneshafrooz, N., Bagherzadeh Cham, M., Majidi, M., & Panahi, B. (2022). Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the WGCNA and LASSO algorithms. *Scientific reports, 12*(1), 20144. [CrossRef]

[3] Faghri, F., Brunn, F., Dadu, A., Chiò, A., Calvo, A., Moglia, C., ... & Traynor, B. J. (2022). Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. *The Lancet Digital Health, 4*(5), e359-e369. [CrossRef]

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

[5] Lusa, L. (2012, December). Evaluation of smote for high-dimensional class-imbalanced microarray data. In *2012 11th international conference on machine learning and applications* (Vol. 2, pp. 89-94). IEEE. [CrossRef]

[6] Grollemund, V., Pradat, P. F., Querin, G., Delbot, F., Le Chat, G., Pradat-Peyre, J. F., & Bede, P. (2019). Machine learning in amyotrophic lateral sclerosis: achievements, pitfalls, and future directions. *Frontiers in neuroscience, 13*, 135. [CrossRef]

[7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21*(9), 1263-1284. [CrossRef]

[8] Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus, 5*(1), 1304. [CrossRef]

[9] Cady, J., Allred, P., Bali, T., Pestronk, A., Goate, A., Miller, T. M., ... & Baloh, R. H. (2015). Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. *Annals of neurology, 77*(1), 100-113. [CrossRef]

[10] Marriott, H., Kabiljo, R., Hunt, G. P., Khleifat, A. A., Jones, A., Troakes, C., ... & Iacoangeli, A. (2023). Unsupervised machine learning identifies distinct ALS molecular subtypes in post-mortem motor cortex and blood expression data. *Acta neuropathologica communications, 11*(1), 208. [CrossRef]

[11] Nguyen, H. L., Vu, D. L., & Le, H. C. (2024, July). Exploiting machine learning and gene expression analysis in amyotrophic lateral sclerosis diagnosis. In *2024 Tenth International Conference on Communications and Electronics (ICCE)* (pp. 363-368). IEEE. [CrossRef]

[12] Rad, H. N., Su, Z., Trinh, A., Newton, M. H., Shamsani, J., Karim, A., ... & Nygc Als Consortium. (2024). Amyotrophic lateral sclerosis diagnosis using machine learning and multi-omic data integration. *Heliyon, 10*(20). [CrossRef]

[13] Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research, 30*(1), 207-210. [CrossRef]

[14] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). Wiley.

[15] Van Rheenen, W., Diekstra, F. P., Harschnitz, O., Westeneng, H. J., van Eijk, K. R., Saris, C. G., ... & van den Berg, L. H. (2018). Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PloS one, 13*(6), e0198874. [CrossRef]

[16] Weeraratne, N., Hunt, L., & Kurz, J. (2024). Challenges of Principal Component Analysis in High-Dimensional Settings when n<p. [CrossRef]

[17] Wang, X., Liu, J., Zhang, Y., Liu, F., & Shen, B. (2009, June). Bioinformatics Analysis of Amyotrophic Lateral Sclerosis Associated Amino Acid Mutations. In *2009 3rd International Conference on Bioinformatics and Biomedical Engineering* (pp. 1-4). IEEE. [CrossRef]

[18] Yang, A., Wang, X., Shang, C., Hu, Y., Yu, C., Zhang, J., & Hong, Y. (2022). Identification of cuproptosis related genes in diagnosis and subtype classification of ALS using the Gene Expression Omnibus Database. [CrossRef]

[19] Tiwari, S., & Shukla, A. (2025). Review on classification of amyotrophic lateral sclerosis using ensemble classifiers. *Engineering Proceedings*, *82*(1), 114. [CrossRef]

[20] Yang, Y., & Ma, G. (2010). Ensemble-based active learning for class imbalance problem. *Journal of Biomedical Science and Engineering*, *3*(10), 1021. [CrossRef]

Eng. Shimaa M. El Makki Osman is postgraduate student, she is a Senior Medical Engineer at SSMO, SUDAN, she received her B.Sc. in Medical Electronics, Faculty of Engineering & Technology, University of Gazira, SUDAN, she is receiving her M.Sc. in Biomedical Engineering from Helwan University, Cairo, Egypt. (Email: shimaaalmkki@gmail.com)

Dr. Esraa M. Hashem She is an Assistant professor in the Biomedical Engineering Department, MUST university, Egypt. Her research interests: Biomedical, Bioinformatics, ML,and Health information systems. She is senior IEEE member and HIMMs member. (Email: Esraa.shebib@must.edu.eg)

Dr. Marwa M. A. Hadhoud is an Associate Professor at the Department of Biomedical Engineering, Faculty of Engineering Helwan University. She received her B.Sc., M.Sc. in biomedical engineering from Faculty of Engineering Helwan University in 2003 and 2007 respectively. In 2012 she received a dual degree Ph.D. in biomedical engineering from Faculty of Engineering Helwan University and Politecnico di Torino Italy. Her research interests are medical signal/image processing, pattern recognition, and bioinformatics. (Email: marwa_hadhoud@h-eng.helwan.edu.eg)

Dr. Vidan Fathi Ghoneim is an Associated Professor in the Biomedical Engineering Department, Faculty of Engineering, Helwan University, Egypt and Princess Nourah Bint Abdulrahman University, Riyadh, KSA. Her research focuses on bioinformatics, digital signal processing, and pattern recognition, with particular interest in medical data analysis. (Email: vfghoneim@pnu.edu.sa; Vidanfathighoneim@h-eng.helwan.edu.eg)