Using Supervised Machine Learning Methods to Create a Gene-Based ALS Predictor from Postmortem Transcriptomics Data

Christopher Bain\textsuperscript{1,2}, Divya Ramamoorthy\textsuperscript{2}, and Ernest Fraenkel\textsuperscript{2}
\textsuperscript{1}Department of Physics and Computer Science, University of Maryland, Baltimore County
\textsuperscript{2}Department of Biological Engineering, Massachusetts Institute of Technology

I. ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is a devastating and progressive neurological disorder that results from the degeneration of motor neurons in the brain and spinal cord. Despite the identification of several pathogenic mutations, the majority of ALS cases are sporadic and lack a clear genetic cause. In this study, we utilized supervised machine learning techniques to analyze transcriptomic data from ALS post-mortem motor cortex samples to identify gene expression patterns that differentiate ALS from healthy controls. Using a logistic regression classifier on a dataset of 276 post-mortem transcriptomic samples collected from multiple academic institutions, we were able to develop a model that predicted ALS with a high degree of accuracy, achieving an F1 score of 90.9%. We also performed decision weight rankings to identify genes that were most strongly associated with the ALS samples, and found that our model accurately identified genes that are currently the subject of active research. Overall, our findings suggest that machine learning techniques can be used to develop accurate predictive models for ALS postmortem tissue, and that these models have the potential to help generate hypotheses for biological pathways that contribute to ALS.

II. INTRODUCTION, PROBLEM & MOTIVATION

Amyotrophic Lateral Sclerosis (ALS) is a fatally progressive, paralytic disorder characterized by the degeneration of motor neurons in the brain and spinal cord. Typically, death due to respiratory paralysis occurs within 3 to 5 years of diagnosis (1). While several pathogenic mutations have been identified, most ALS cases have no family history of disease and can be a product of multiple pathways contributing to varying degrees in each patient (2).

Motor neurons are organized into upper and lower populations in the motor cortex, brain stem, and spinal cord. Upper motor neuron failure results in muscle stiffness, while lower motor neuron failure leads to spontaneous muscle twitching and atrophy as they lose synaptic connectivity with their target muscles. ALS is difficult to diagnose early because it can mimic other neurological diseases until the late stages of diseases, making diagnosis challenging until symptoms have progressed (1, 2, 3). Further, there are currently a limited number of effective disease-modifying therapies or cures for ALS (3), and thus, better understanding the biological mechanism behind the disease is critical (2, 4).

In recent years, advances in machine learning techniques have allowed for the analysis of large-scale transcriptomic and proteomic data (8, 9, 10), offering new insights into the molecular mechanisms underlying ALS pathogenesis (2, 7). In this study, we utilized supervised machine learning algorithms to analyze post-mortem transcriptomic data from multiple academic institutions (Barrow Neurological Institute, Columbia University Medical Center, Georgetown University, Johns Hopkins University, Mount Sinai Hospital, Rutgers University in collaboration with the NIH, University College London, University of California San Diego, and University of Maryland) and developed a predictive model that accurately identifies ALS postmortem samples. Our study aims to contribute to the characterization of the biological pathways underlying the disease.

Supervised machine learning is a powerful technique that can be used to develop predictive models from labeled datasets (11, 12). In supervised learning, an algorithm is trained on a set of inputs (features) and their corresponding outputs (labels), with the goal of learning a general mapping from inputs to outputs. Once trained, the model can be used to predict the output for new, unseen inputs.

In the context of ALS research, supervised machine learning
Identifying ALS with Supervised Methods

Out of 146,765 Transcripts with Nonzero Total Count

<table>
<thead>
<tr>
<th></th>
<th>Adjusted p-value &lt; 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFC &gt; 0 (up)</td>
<td>380, 0.26%</td>
</tr>
<tr>
<td>LFC &lt; 0 (down)</td>
<td>321, 0.22%</td>
</tr>
<tr>
<td>Outliers</td>
<td>853, 0.58%</td>
</tr>
<tr>
<td>Low Counts</td>
<td>50508, 34%</td>
</tr>
</tbody>
</table>

(mean count < 4)

Table I: Statistical results breakdown of differentially expressed genes between a pilot of 6 patient samples generated using DESeq2.

learning can be used to identify relevant biomarkers that are associated with the disease ([13], [14]). For example, a study might collect data on gene expression levels, protein levels, or neuroimaging features from both ALS patients and healthy controls, and use supervised learning to identify which features are most predictive of ALS status [14]. Once these features are identified, they can be used to develop predictive tools that can aid in the diagnosis or prognosis of the disease. In our case, we used this classifier to take in the gene transcriptomics from postmortem motor cortex samples and produce the binary result of ALS or Control.

III. METHODOLOGY

A. DESeq2 Analysis

In this study, we analyzed the comprehensive dataset of 335 patients, covering 146,765 transcripts obtained from Tam et al. 2019 [2]. Brain samples were collected from either the medial or lateral cortices, with nearly equal representation from both regions, comprising 49.9% and 50.1% of the dataset, respectively. Gender distribution was also accounted for, with 56.7% of the dataset consisting of male samples and 43.3% consisting of female samples. To focus on understanding differences between ALS and healthy controls, the dataset was pruned to 276 patients, with a split of 75.4% ALS and 24.6% control subjects.

To be able to be confident in our ability to gauge if a machine learning classifier could make accurate determinations of a tissue sample being derived from an ALS patient, we had to first analyze if there were enough differentially expressed genes between the ALS and control group samples. To do this we implemented the use of the DESeq2 [15] library in R to find up and down regulated genes between the groups.

As a pilot analysis, we employed a sample size of six patients from our transcriptomic database and performed statistical analysis using DESeq2 to identify the differential expression of genes. Specifically, our aim was to discern the prevalence of up-regulated and down-regulated genes in our dataset, with the ultimate goal of gaining a deeper understanding of the differences between the genes of ALS patients and healthy controls.

Our analysis, as seen in Table I, revealed a substantial number of differentially expressed genes, both up-regulated and down-regulated, within and between the studied samples. These results provided a solid foundation for us to proceed with machine learning-based classification.

B. Supervised Learning Classification

We next approached our task of understanding ALS postmortem gene expression patterns using the logistic regression classifier provided by the SKLearn library in Python [16]. This particular machine learning algorithm was selected for its proficiency in processing feature data and yielding a binary outcome ([12], [14]). To train the model, we allocated 70% of the available data to the training set and the remaining 30% to the test set. This approach is commonly used in the machine learning field and helps to prevent overfitting by evaluating the model’s performance on previously unseen data. Ultimately, we evaluated the performance of our model on a test set cohort of 83 patients.

We included all genes as input in our classifier to better understand their combinatorial effects on classification performance, using kallisto count abundances [17] and normalized by transcripts per million (tpm) for all available transcripts. Kallisto pseudoaligns RNA-seq reads to a reference without aligning individual bases, producing a list of compatible transcripts. It achieves similar accuracy to previous approaches but is significantly faster, enabling the analysis of large amounts of RNA-seq data in a fraction of the time [17].

After splitting the data into training and test sets, we proceeded to train the logistic regression classifier on the training set. We utilized default training parameters from the sklearn.linear_model.LogisticRegression classifier to allow the model to self-optimize parameters for decision-making. We then trained the logistic regression classifier on the entire training set and then applied it to the test set to evaluate its performance. We used
standard evaluation metrics to assess the classifier’s performance on the test set. Additionally, we generated a receiver operating characteristic (ROC) curve for the test set to visually evaluate the tradeoff between true positive rate and false positive rate at different classification thresholds.

C. Uniqueness of Approach

Unlike traditional approaches that rely on laboratory experimentation, this study aims to explore the potential of predicting ALS using transcriptomic data and classification algorithms. The innovative nature of this approach stems from its potential to circumvent the limitations of conventional methods, which are often time-consuming, labor-intensive, and costly. By leveraging the power of supervised machine learning techniques, this study seeks to establish a model for predicting ALS postmortem samples, which could ultimately pave the way for understanding the mechanism for this debilitating disease.

IV. RESULTS

A. Classification Accuracy

Following the data processing phase, the model demonstrated a high degree of accuracy in its ability to discern between postmortem ALS and control tissue. Specifically, our analysis yielded accuracy and F1 scores of 86.7% and 90.9%, respectively, indicating that the model performed well in its classification task (Figure 3a).

Moreover, the model’s performance was analyzed on a ROC curve, which showed that it produced accurate results. The area under the curve (AUC) score of 0.947 further highlights the model’s effectiveness in accurately classifying postmortem cortex samples (Figure 3b).

These results provide evidence for the potential of machine learning in ALS research. With the use of a simple machine learning model and gene expression data, we were able to accurately differentiate between ALS and control postmortem tissue samples. The findings of this study suggest that machine learning algorithms could be utilized to understand the underlying biomechanics of ALS.

B. Ranking of Genes by Feature Weighting Coefficients

To gain insight into how the model was making these predictions, we examined the absolute values of the coefficients for the top 10 specific features used across the entire population of training data. This analysis allowed us to identify the genes that played a more important role in the classification of ALS in the postmortem tissue samples.

Upon ranking the genes, we found that CYTB was 100% more weighted than all other genes ranked within the top 10 of the list (Figure 4). Our findings suggest that CYTB, along with ND4, COX3, and COX1, contribute strongly to the model’s overall ability to predict a correct classification of ALS in postmortem motor cortex tissue. Further examination of these genes revealed that they have previously been identified as potential contributors to mitochondrial death ([18], [19], [20]), inflammatory response ([21], [22], [23]), and muscle-weakening of patients throughout the ALS life cycle ([24], [25], [26]).

It is worth noting that our model had no access to available literature on ALS and the genes particularly being studied in the field at the time. Despite this, the model was able to suggest potential biomarkers for more substantial prediction of the disease. These results suggest that postmortem tissue may provide valuable insight into the identification of potential biomarkers for ALS, which may aid in better characterization of the molecular mechanisms of the disease and aid in searching for disease treatments.

V. DISCUSSION

Our analysis revealed that the mitochondrial protein CYTB played an influential role in classifying whether a postmortem sample was derived from an ALS patient or healthy control. Moreover, other genes such as ND4, COX3, and COX1 were also found to contribute significantly to the overall determination of a result for the model (Figure 4). While the findings of our study are promising, several limitations must be considered.

First, the sample size used in our analysis was relatively small, comprising only 208 ALS patients and 68 controls. Therefore, future studies should aim to expand the sample size to evaluate the robustness of these findings. In addition, further investigation is required to determine whether the genes identified in this study have functional relevance in the pathogenesis of ALS.

It is also crucial to acknowledge that the exact role of the molecules identified by the model in the context of ALS remains unclear, and further research is required to better understand their functional relevance. In par-
Figure 3: a) Confusion matrix generated on test set (83 samples) to evaluate the predictive performance of the Logistic Regression classifier. Notably, the number of true positive and true negative classifications constitute a majority of the model’s predictions, indicating high accuracy in identifying differentiating between samples. b) Receiver Operating Characteristic (ROC) curve generated for the Logistic Regression classifier. The curve illustrates the performance of the model, showing strong discriminatory power between ALS and control groups.

Figure 4: Top 10 genes, ranked based on the absolute value of coefficients assigned by the Logistic Regression classifier, with representative scores. Notably, Cytochrome B (CYTB) exhibits a substantially higher score than the other genes featured on the list.

ticular, it is difficult to differentiate causal from compensatory signal when data is derived from postmortem tissue alone. It is also important to note that the identification of molecules that have been studied in the context of ALS does not prove that they are biomarkers of the disease. Hence, further studies are needed to investigate the role of these genes in ALS, as well as to examine the efficacy of machine learning approaches in predicting ALS biomarkers at earlier stages of the disease.

A. Future Work

In this study, we have successfully identified genes that demonstrate a high correlation with the presence of ALS in postmortem tissue. However, the significance and functional relevance of these genes in the pathogenesis of ALS still require further investigation. The identification of potential biomarkers and targets for ALS treatment is of great significance in advancing the field of biomedical engineering and the development of new therapeutic approaches. Future studies will be required to investigate if comparable signals can be identified from easily accessible samples such as blood, as the samples utilized in this study were postmortem. Future studies should aim to elucidate the molecular mechanisms underlying the contribution of these genes to the pathogenesis of ALS, whether they act as causative factors or compensatory mechanisms. Understanding the role of these genes in the disease progression of ALS is critical in advancing the understanding of the disease and ultimately developing new treatments to improve patient outcomes.

A potential avenue of research involves applying unsupervised machine learning methods to the dataset in question. While supervised learning methods have shown promise in classifying postmortem ALS tissue samples, they require class labels, in this case ALS versus healthy control, to be pre-specified. By contrast, unsupervised learning methods do not require predetermined class labels. This can allow researchers to begin to understand the underlying structure of the data they are working with, similar to methods used by Tam et al. 2019.
Further, the use of unsupervised learning in this context may be helpful for identifying subgroups of ALS, such as how ALS can be caused by different pathologies. A better understanding and identification of these subgroups could prove helpful in developing more effective therapies. Utilizing unsupervised learning techniques to identify patterns and relationships in the data could provide new insights into the underlying mechanisms of ALS and inform the development of more effective treatments. However, it is important to note that unsupervised learning methods can be difficult to evaluate and require additional validation steps to be applied to learned subgroups.

B. Contributions and Impacts

The research discussed in this paper provides valuable insights into the development of machine learning models for identifying potential biomarkers for Amyotrophic Lateral Sclerosis (ALS), a progressive neurodegenerative disease with no known cure. By utilizing a logistic regression classifier on a dataset of gene expression profiles from ALS motor cortex samples, the study identifies genes that play an influential role in the classification of ALS from healthy controls in postmortem tissue, most notably the mitochondrial protein CYTB. The identification of these genes and their potential functional relevance in the pathogenesis of ALS provides new avenues for understanding and potentially treating this devastating disease.

Furthermore, the study demonstrates the potential for machine learning methods in identifying biomarkers and understanding disease progression. As machine learning methods become more sophisticated and accessible, they have the potential to revolutionize the field of medicine and provide new approaches for diagnosis, treatment, and personalized medicine. The broader impact of this research lies in its potential to not only advance our understanding of ALS, but also to pave the way for future developments in the field of computational biology and biomedical engineering, with the ultimate goal of improving healthcare outcomes for patients.