An Interpretable Framework for single nucleotide polymorphism-based Amyotrophic Lateral Sclerosis risk prediction through supervised machine learning

Shaila Malkani

Introduction

Common and complex non-communicable diseases, such as neurodegenerative disorders, cancers, and autoimmune diseases are pressing challenges facing modern-day healthcare. Early detection of these diseases plays a pivotal role in their effective prognosis and prevention. Up until the last decade, risk prediction for susceptibility towards certain diseases heavily relied on traditional epidemiological measures in clinical settings, wherein factors such as lifestyle, family history, and patient demographics, etc. were taken into account. However, these models have proven to have limited predictive power and contained inaccuracies, because the other half of the equation was missing.Unlike rare (monogenic) diseases that follow a clear Mendelian pattern of inheritance as a result of the dysfunction of single genes, common diseases follow a polygenic inheritance pathway which makes risk prediction more challenging. Since the completion of the Human Genome Project, there has been an exponential increase in the availability and efficiency of genome sequencing technology, as well as a significant decrease in the cost per base pair sequencing (Ho et al.). As a result of the advancing technology with reducing costs, the idea of using genetic variants (such as Single Nucleotide Polymorphisms) for genome-based risk prediction has rapidly transformed the way researchers approach the detection of diseases and has had a myriad of benefits on disease prevention. The use of accurate risk prediction eliminates the need for expensive and invasive screening methods to predict the onset of a disease.

Currently, the primary method of conducting genome-based risk prediction for polygenic diseases has been polygenic risk scoring (PRS). A PRS measures the effect that genetic variants associated with a trait affect an individual's likelihood of developing it. The SNP variation used to calculate polygenic risk scores is primarily derived from genome-wide association studies (GWAS). Essentially, GWAS have successfully been able to map genetic variants from disease-associated loci to specific polygenic disorders (Hirschhorn and Daly). These associations are constructed by analyzing genetic variants that have a higher frequency in individuals with the disorder (cases) than in the rest of the healthy population (controls) (Cano-Gamez and Trynka). GWAS have played a pivotal role in understanding the underlying molecular mechanisms behind complex diseases and have helped demystify the role that genetic variation plays in causing disease.

Recently, machine learning methods have leveraged SNP data from GWAS to produce powerful computational algorithms for accurate risk prediction (Cano-Gamez and Trynka). These methods have the potential to statistically map SNP variation from genotypes to the onset of complex disease phenotypes. Risk prediction as a classification problem involves the use of

SNP minor allele frequency as the input data (features) and the output as the diagnosis of the specific disease (classification label). Supervised machine learning methods have shown promise to change the face of risk prediction, while also giving us notable insight into the polygenic architecture of common diseases faced by humans today.

One of these complex diseases that could greatly benefit from the intervention of machine learning is Amyotrophic lateral sclerosis (ALS). ALS is a chronic neurodegenerative disorder that affects around one in 50,000 people per year. It is characterized by a progressive loss of muscle control, leading to speaking, eating, swallowing, and breathing difficulties (Heller). Unfortunately, the etiology of ALS remains to be unknown and therefore there is no cure for this fatal disease (Heller). However, supervised machine learning shows great promise to be able to predict an early diagnosis of ALS based on SNPs that an individual may possess (Genet). The use of machine learning would pose a myriad of benefits for patients who could potentially develop the chronic disease, and it would also give us essential insights into the molecular mechanisms behind ALS as well as potential therapeutics that could improve the quality of life for these individuals.

Previous work

Recently, supervised machine learning approaches have emerged as powerful tools for genome-based risk prediction for various different common diseases such as cardiovascular disease, Alzheimer's disease, cancers, and type 2 diabetes. While the use of SNPs for risk prediction is still a relatively new area, previous literature has shown that several genome-based risk prediction algorithms have been more successful than standard PRS.

In a recent study, Gaudillo et al. were able to achieve an area under the receiver operator characteristic curve (AUC) score of 0.62 for asthma risk prediction using an integrated Random Forest and SVM model, and a score of 0.62 for an integrated Random Forest and k-nearest neighbors model (Gaudillo et al.). SNP data has also been used to investigate whether steroid-metabolism gene SNPs would cause breast cancer due to an increase of hormones and environmental toxins. Dumitrescu and Cotarla were able to find an optimal model by testing whether SVM, Naive Bayes, or decision tree classifiers worked best on the data, and by reducing their sample size from 98 SNPs to 2-3 SNPs. The SVM classifier was able to achieve the highest accuracy of 69% (Dumitrescu and Cotarla). These studies highlight how machine learning approaches for genome-based risk prediction have promising potential to be clinically useful and give researchers insight into the molecular mechanisms that lead to disease.

In the case of ALS, Gupta et al. have utilized machine learning for risk prediction by measuring epidemiological factors such as patient environment and family history (Gupta et al.). However, this study did not use SNP data as one of the variables, which is a key factor that could perhaps play a role in how susceptible an individual is to developing ALS. Currently, there have been no machine learning methods for ALS risk prediction informed by SNP data, which elucidates the need for this investigation, as it could potentially add another dimension to the risk prediction of ALS. Through this investigation, we aim to explore how supervised learning methods can be

used to predict the diagnosis of ALS based on single nucleotide polymorphism minor allele counts.

Methods

Data acquisition and cleaning

The dataset chosen for this project was genomics and clinical data from the End ALS Kaggle Challenge ("End ALS Kaggle Challenge"). The genomics data consists of SNP information along with header information such as the variant chromosome, the variant position within the chromosome, the variant ID (rsID from dbSNP), as well as the reference and alternate allele bases from 134 people with ALS or motor neuron disease, and a control population of 30 people without the diseases. All of the variants listed in the dataset are also filtered through the GATK VQSR quality control, eliminating variants that are present in intergenic regions and have a frequency higher than 10%. The variants are also coded in binary data, in which patient genotypes that only have the reference allele are labeled as 0, whereas the genotypes with one or both copies of the alternate allele are labeled as 1.

Additionally, the second dataset used for the project was an ALS GWAS summary statistics dataset acquired from Project MinE (Genet). This dataset contains cross-ancestry GWAS data from 29,612 ALS patients (cases) and 122,656 healthy individuals (controls). The summary statistics contain information on the chromosome number, SNP rsID, alternate and reference allele bases, frequency, as well as the p-value for each SNP. The p-value is the probability that the specific SNP is associated with the disease of interest. In this case, SNPs with a lower p-value are statistically more associated with developing ALS.

We then pre-processed and analyzed for machine learning, primarily using built-in Python libraries such as Pandas, and numPy. The first step in this process was to align the genomics data frame containing SNP data corresponding to each participant ID to the clinical data frame containing the participant IDs that aligned to whether the patient was a case (with ALS) '1' or control (healthy) '0'. We then sorted the SNPs from the summary statistic dataset in order of ascending p-values, and the top 100 SNPs that were present in the summary statistics as well as the genomics data frame were chosen as the features for the machine learning model. The labels for the machine learning model were from the clinical data frame, wherein a predicted case of ALS would result in '1', and the absence of ALS would correspond to a value of '0'.

Machine learning models

The supervised machine learning models applied to the data were logistic regression, random forest, and Naive Bayes using the Python library Scikit-learn (scikit-learn). The initial goal of the project was to establish benchmark models using default parameters. We then aimed to make improvements to the model by experimenting with class balance, optimization of hyperparameters, and regularization techniques.

Logistic Regression

Logistic regression is a supervised classification algorithm used to produce discrete outcomes in a binary manner when given input variables (Edgar and Manz). In this case, the input variables (or features) are the top SNPs and the output variable is the diagnosis of ALS (0 or 1). The model learns the relationship between the variables from the labeled dataset in the form of a sigmoid function. The advantage of using logistic regression as the baseline model is that it is relatively easy to implement, train and interpret. Additionally, it has the ability to produce a measure of the direction of association between the independent and dependant variables, which in this case is whether the SNPs have a positive or negative effect on the predicted diagnosis of ALS. However, an important drawback of logistic regression is that it is highly susceptible to overfitting in high-dimensional datasets (Rout). Because logistic regression is a model that tends to overfit data, it is essential that different regularization techniques are used to combat the likelihood of overfitting to improve the model performance. Another drawback of using logistic regression on this dataset is that the model assumes the absence of multicollinearity between the features (Stoltzfus). However, due to linkage disequilibrium, many SNPs that are analyzed through GWAS are often highly correlated with one another. Linkage disequilibrium occurs when there is a correlation between nearby SNPs at different loci on the same chromosome. Variants are in linkage disequilibrium when the frequency of association is higher than if they were unlinked or associated randomly (Slatkin).

Random Forest

Random forest is another supervised classification algorithm that is constructed from multiple decision trees. The algorithm establishes the outcome based on the predictions made by individual decision trees, and makes predictions by calculating the mean of the outputs from the decision trees. The decision trees represent separate and distinct classification instances of the inputted data, and the random forest selects predictions based on the majority of votes (“Random Forest - Overview, Modeling Predictions, Advantages”). The primary advantage of random forest is that it performs well on high-dimensional data with numerous features, which is a feature that is highly beneficial for this dataset. Additionally, random forest can also balance any imbalanced classes automatically, which is why it is suitable to use on the imbalanced dataset used in this investigation.

Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes theorem, and carries a strong assumption of independence between predictor variables. Essentially, Naive Bayes works on the concept that each predictor variable has an equal and independent contribution to the final outcome (Rish). While this assumption is beneficial for allowing the model to perform well on less training data such as the dataset used in this investigation, the independence between features negates the additive effects and interactions between each SNP that may further contribute to the development of ALS. As mentioned previously, it is likely that most SNPs will be correlated with each other due to linkage disequilibrium but Naive Bayes assumes that the associations between each SNP do not exist, which in theory would have an effect on the performance of the model. However, studies have shown that in practice, the independence

assumption that Naive Bayes makes allows it to compete with other sophisticated classifiers (Rish).

Model training

For each model, we split the dataset into training and testing sets, using an 80:20 ratio. This meant that 80% of the data was assigned to training the model during the development stage, whereas 20% was allocated to the testing set where the trained models were used to make predictions on the unseen data. This was to ensure that the models were able to generalize the learned relationship to new data from the test set. After the splitting of the datasets using Scikit-learn, each classification model was imported, initialized, and fitted to the training data without tuning any hyperparameters to benchmark the baseline models. The trained models were then used to make predictions on the test set and evaluated based on performance.

Evaluating baseline performance

The performance of the baseline models was evaluated using accuracy as well as the area under the receiver operating characteristics curve (AUROC). Accuracy is a commonly used metric to evaluate performance. Accuracy is calculated by comparing the number of correctly predicted samples to the total number of samples, and follows a scale of 0 (absence of correct predictions) to 1 (every prediction is correct) (Zvornicanin). However, because the dataset is highly imbalanced, accuracy can be a misleading measure of performance. AUROC is another commonly used metric when dealing with data that contains class imbalance as it uses prediction probabilities (“Measuring Performance: AUC (AUROC)”). The receiver operating characteristics curve (ROC) essentially outlines the relationship between true positive rate and false-positive rate for different probability thresholds. AUROC is used to measure the model’s ability to discriminate between positive and negative cases, and follows a scale of 0 (absence of correct predictions) to 1 (all predictions are correct). An AUROC value of 0.5 indicates that the model is either predicting random or constant class points (Zvornicanin). Additionally, the area under the precision-recall curve (AUPR) is another performance metric similar to AUROC, but it is used on highly imbalanced datasets as it focuses on the fraction of true positives in the positive classes rather than the negative classes (Saito and Rehmsmeier). The AUPR is also the average of the precision scores which is calculated at different thresholds.

Initial results

Baseline models

We trained baseline models with default parameters using random forest, logistic regression, and Naive Bayes. The AUROC performances for the baseline random forest, logistic regression, and naive Bayes models were 0.49, 0.43, and 0.48 respectively. These results show that the baseline models performed worse than random, and therefore need to be improved using various methods in order to be useful in a clinical setting.

Dealing with imbalanced classes

The first improvement we made to the models was balancing the class weights within the training dataset. This was an essential step because the dataset used is highly imbalanced, as the case genotypes (134 individuals) are significantly larger and overrepresented than the control genotypes (10 individuals). This is an important issue because it is likely that the machine learning models were able to predict and be biased towards the over-represented positive instances (ALS diagnosis) rather than the scarce negative instances (absence of ALS). This issue can potentially be combatted through different approaches, however, the most feasible one (considering the size of the dataset) was through balancing the class weights (Singh).

To solve this challenge, more weight can be given to the minority class (without ALS) so that the algorithm focuses on reducing the errors within the minority class, without letting it become biased towards this class instead. A built-in parameter within Scikit-learn was used to balance the classes, by automatically assigning class weights that are inversely proportional to the frequency of each class (Singh). Although this process helps optimize the class imbalance problem, the drawback is that it increases true recall at the cost of decreasing true precision ("Using Class Weight to Improve Class Imbalance").

Optimization of hyperparameters

The next improvement made to the models was optimizing the parameters to improve model performance. For each model, a different set of hyperparameters were tuned both manually, as well as using a grid and random search. A grid search is conducted by defining a grid of hyperparameters and evaluating every position in the grid using cross-validation. The most optimal parameters from the grid search are defined as the point of the grid that maximizes the performance in cross-validation. On the other hand, random search only tests a subset of the points on the grid rather than the entire grid to find the best hyperparameters. For example, a grid search on the Naive Bayes model showed that by tuning the var_smoothing hyperparameter, the testing AUROC value of the model increased from 0.710 to 0.843, which was an 18.3% increase in performance.

Additionally, it is important to consider the overfitting of the models to the training data. All three models originally presented a high training AUROC value of 1.0, whereas the testing AUROC values were significantly lower. The disparity between training and testing AUROC values was likely a result of overfitting, wherein the models tend to memorize the training data instead of learning the relationship between the variables, resulting in poor performance on unseen data (Ying). To combat overfitting, C, L1, and elastic-net regularization techniques were used to try to overcome the challenge of the overfit models. For example, the grid search on the logistic regression model showed that a C value of 1×10^{9} was beneficial in increasing the testing AUROC value from 0.43 to 0.56.

Feature importances

We calculated the importance of the features for each model in order to compare them to the p-value for each SNP. For logistic regression, this process involved examining the model

coefficients. As mentioned earlier, the advantage of logistic regression is that it stores the coefficients assigned to each feature as well as the direction of the effect, demonstrated through the positive and negative coefficients associated with each feature ("Scikit-Learn 0.21.2 Documentation"). When the feature importances were calculated for each model, we found that the feature importances did not correspond to the p-value of the association between each SNP and the disease. The high feature importance could perhaps be attributed to the interactions between the SNPs rather than singular SNPs having a large effect on developing the disease.

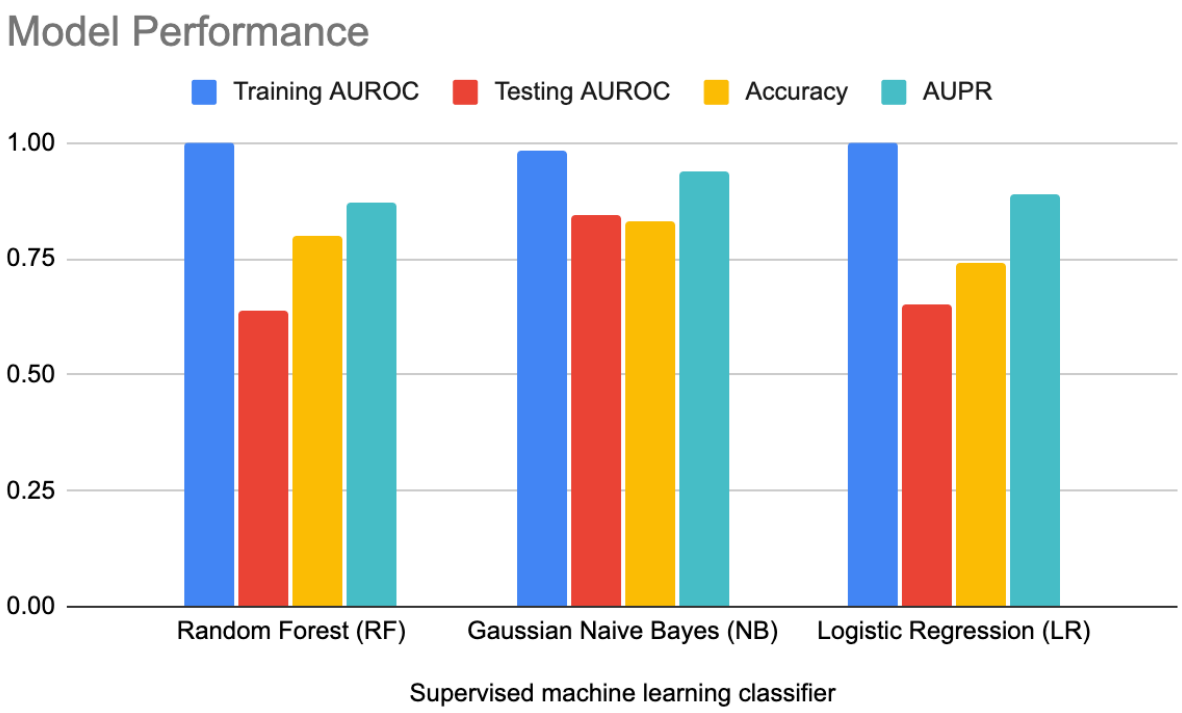
Final results

After improving the baseline models through optimization of hyperparameters, using regularization techniques, and balancing class weights, we trained the final random forest, Naive Bayes, and logistic regression models. We evaluated the performance of these models using metrics such as training and testing AUROC, accuracy, and AUPR as shown in table 1. The results illustrate that while the random forest classifier and logistic regression had training AUROC scores of 1.0, Naive Bayes had the highest testing AUROC of 0.843. The highest accuracy of 83% also belonged to Naive Bayes, followed by random forest with an accuracy score of 80%, and logistic regression with the lowest accuracy of 74%. Figure 1 outlines the comparison between AUROC, accuracy, and AUPR of each model.

Table 1. Summary of model performances of random forest, Naive Bayes, and logistic regression classifiers. The metrics we used to evaluate performance include training and testing AUROC, accuracy, and AUPR.

Supervised machine learning classifier	Training AUROC	Testing AUROC	Accuracy	AUPR
Random Forest (RF)	**1.000**	0.640	80%	0.87
Gaussian Naive Bayes (NB)	0.985	**0.843**	**83%**	**0.94**
Logistic Regression (LR)	**1.000**	0.653	74%	0.89

Figure 1. Comparison of final model performances of random forest, Naive Bayes, and logistic regression classifiers.

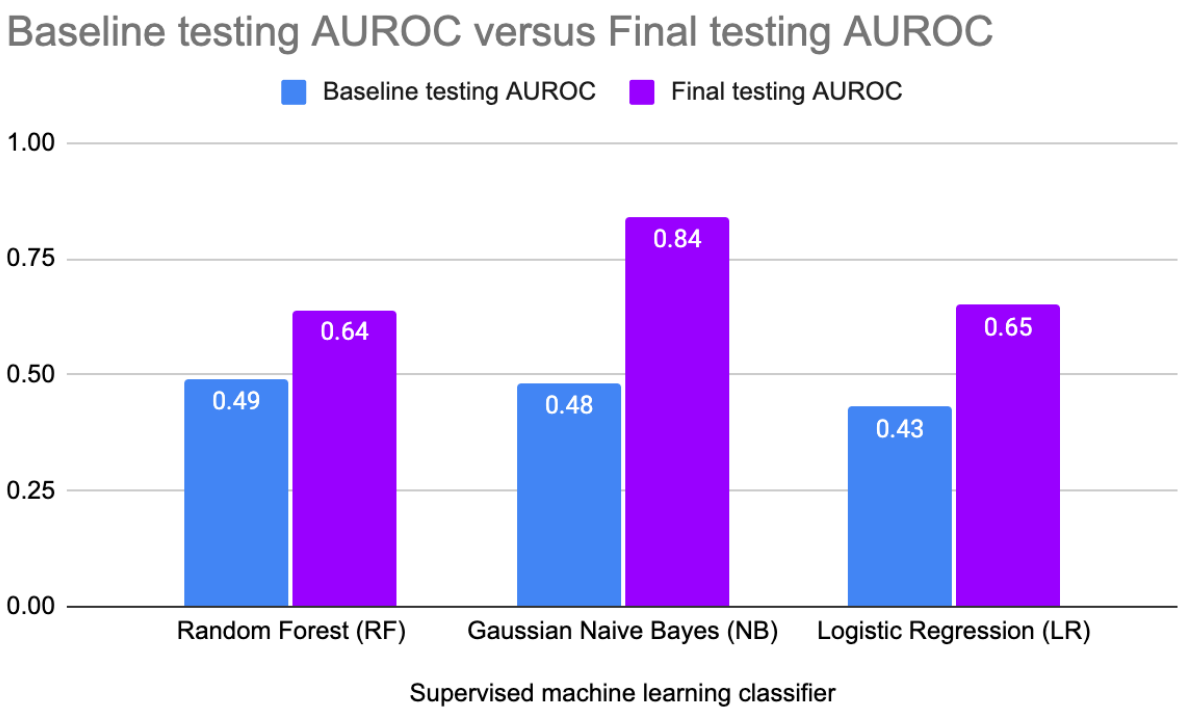


Additionally, Table 2 and Figure 2 highlight the comparison between the AUROC score of the baseline models and the AUROC score of the final models. The results show that after the improvements to the models were made through regularization, dealing with the class imbalance and hyperparameter optimization,

Table 2. Summary of baseline model AUROC performance and final model AUROC performance

Supervised machine learning classifier	Baseline testing AUROC	Final testing AUROC
Random Forest (RF)	**0.49**	0.64
Gaussian Naive Bayes (NB)	0.48	**0.84**
Logistic Regression (LR)	0.43	0.65

Figure 2. Comparison between baseline model AUROC performance and final model AUROC performance



As shown in figure 2, the Naive Bayes classifier outperformed random forest and logistic regression based on final testing AUROC, and also demonstrated the greatest increase in performance from an AUROC score of 0.48 to 0.84. Therefore, the Naive Bayes classifier has shown to be the most optimal classifier for this dataset, followed by logistic regression and random forest.

Discussion

This investigation demonstrates how supervised machine learning can be used as a powerful tool for SNP- based risk prediction for ALS. As highlighted earlier, ALS is a fatal disease that could greatly benefit from the intervention of machine learning for risk prediction, and therefore, early detection. Currently, ALS is most commonly diagnosed after individuals have already suffered irreversible damage to motor neurons present within the central nervous system. Studies have shown that the early detection of ALS could not only delay and slow down the disease progression, but also has the potential to even prevent extensive neuron loss as a result of the disease (“New Program Hopes to Make Early Detection and Treatment of ALS a Reality”).

However, the results from this investigation pose certain limitations on the scope of our models, and whether they can be potentially beneficial in clinical settings. Although the highest accuracy score from our analysis was 83% for the final Naive Bayes model, it is important to consider that

a large margin of error can be particularly detrimental in clinical settings. For example, in the case of false-negative results, individuals would be unaware that they are susceptible to a particular disease based on their genetic variation and may not take any preventative measures and eventually develop the disease. In this case, one could argue that a false-negative result is more dangerous than a false-positive result because extraneous factors such as environment and family history also play an important role in whether an individual develops a disease or not, and therefore it is best to take preventive measures either way. Because of this, it is critical to emphasize the idea that risk-prediction models should not be used as diagnostic tools, but should instead be used to aid medical systems and doctors for early detection and to take preventative measures for patients at risk of developing diseases.

In the future, it is important that we harness the prospect of machine learning not only for ALS risk prediction, but also to gain insights into the polygenic architecture and molecular mechanisms behind ALS, such as the additive effects and epistatic interactions between SNPs that may affect the likelihood of developing the disease. Though this is beyond the scope of this paper, SNP data from GWAS combined with machine learning approaches can also be used to analyze individual responses to potential therapeutics that could improve the quality of life for these ALS patients.

This investigation highlights the potential of supervised machine learning approaches to SNP-based risk prediction. However, due to the scarcity and imbalance within the dataset used to train our models, future work would require larger, unbiased, and more robust data to accurately assess the role of machine learning techniques for ALS risk prediction. In conclusion, the use of machine learning to inform the process of the early detection of ALS in clinical settings can pose countless benefits for individuals who are predisposed to developing the disease, and could potentially save thousands of lives.

References

Cano-Gamez, Eddie, and Gosia Trynka. “From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases.” *Frontiers in Genetics*, vol. 11, 13 May 2020, 10.3389/fgene.2020.00424.

Dumitrescu, R. G., and I. Cotarla. “Understanding Breast Cancer Risk - Where Do We Stand in 2005?” *Journal of Cellular and Molecular Medicine*, vol. 9, no. 1, Jan. 2005, pp. 208–221, 10.1111/j.1582-4934.2005.tb00350.x.

Edgar, Thomas W., and David O. Manz. “Chapter 4 - Exploratory Study.” *ScienceDirect*, Syngress, 1 Jan. 2017, www.sciencedirect.com/science/article/pii/B9780128053492000042.

"End ALS Kaggle Challenge." *Www.kaggle.com*, www.kaggle.com/datasets/alsgroup/end-als. Accessed 23 May 2022.

Gaudillo, Joverlyn, et al. "Machine Learning Approach to Single Nucleotide Polymorphism-Based Asthma Prediction." *PLOS ONE*, vol. 14, no. 12, 4 Dec. 2019, p. e0225574, 10.1371/journal.pone.0225574. Accessed 19 May 2022.

Genet, Eur J Hum. "Project MinE: Study Design and Pilot Analyses of a Large-Scale Whole-Genome Sequencing Study in Amyotrophic Lateral Sclerosis." *European Journal of Human Genetics*, vol. 26, no. 10, 28 June 2018, pp. 1537–1546, 10.1038/s41431-018-0177-4. Accessed 3 Nov. 2021.

Gupta, P.K., et al. "A Predictive Model for Amyotrophic Lateral Sclerosis (ALS) Diagnosis." *Journal of the Neurological Sciences*, vol. 312, no. 1-2, Jan. 2012, pp. 68–72, 10.1016/j.jns.2011.08.021. Accessed 13 Jan. 2022.

Heller, Laura. "ALS, Amyotrophic Lateral Sclerosis, Lou Gehrig's Disease." *Www.hopkinsmedicine.org*, www.hopkinsmedicine.org/neurology_neurosurgery/centers_clinics/als/conditions/als_amyotrophic_lateral_sclerosis.html#:~:text=ALS%20Statistics&text=It%20affects%20as%20many%20as.

Hirschhorn, Joel N., and Mark J. Daly. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics*, vol. 6, no. 2, Feb. 2005, pp. 95–108, 10.1038/nrg1521.

Ho, Daniel Sik Wai, et al. "Machine Learning SNP Based Prediction for Precision Medicine." *Frontiers in Genetics*, vol. 10, 27 Mar. 2019,

www.frontiersin.org/articles/10.3389/fgene.2019.00267/full, 10.3389/fgene.2019.00267. Accessed 25 May 2019.

"Measuring Performance: AUC (AUROC)." *Glass Box*, 23 Feb. 2019, glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/#:~:text=AUROC%20is%20thus%20a%20performance. Accessed 23 May 2022.

"New Program Hopes to Make Early Detection and Treatment of ALS a Reality." *Columbia University Irving Medical Center*, 19 Nov. 2018, www.cuimc.columbia.edu/news/new-program-hopes-make-early-detection-and-treatment-als-reality#:~:text=If%20ALS%20could%20be%20detected. Accessed 23 May 2022.

"Random Forest - Overview, Modeling Predictions, Advantages." *Corporate Finance Institute*, corporatefinanceinstitute.com/resources/knowledge/other/random-forest/.

Rish, Irina. "(PDF) an Empirical Study of the Naïve Bayes Classifier." *ResearchGate*, www.researchgate.net/publication/228845263_An_Empirical_Study_of_the_Naive_Bayes_Classifier.

Rout, Amiya Ranjan. "Advantages and Disadvantages of Logistic Regression." *GeeksforGeeks*, 25 Aug. 2020, www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/.

Saito, Takaya, and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, vol. 10, no. 3, 4 Mar. 2015, p. e0118432, 10.1371/journal.pone.0118432.

scikit-learn. "Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.20.3 Documentation." *Scikit-Learn.org*, 2019, scikit-learn.org/stable/.

Singh, Kamaldeep. “How to Improve Class Imbalance Using Class Weights in Machine Learning.” *Analytics Vidhya*, 6 Oct. 2020, www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/.

“Sklearn.linear_model.LogisticRegression — Scikit-Learn 0.21.2 Documentation.” *Scikit-Learn.org*, 2014, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Slatkin, Montgomery. “Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future.” *Nature Reviews Genetics*, vol. 9, no. 6, June 2008, pp. 477–485, 10.1038/nrg2361.

Stoltzfus, Jill C. “Logistic Regression: A Brief Primer.” *Academic Emergency Medicine*, vol. 18, no. 10, Oct. 2011, pp. 1099–1104, 10.1111/j.1553-2712.2011.01185.x. Accessed 26 Apr. 2019.

“Using Class Weight to Improve Class Imbalance.” *Develop Paper*, 7 May 2021, developpaper.com/using-class-weight-to-improve-class-imbalance/. Accessed 23 May 2022.

Ying, Xue. “(PDF) an Overview of Overfitting and Its Solutions.” *ResearchGate*, www.researchgate.net/publication/331677125_An_Overview_of_Overfitting_and_its_Solutions.

Zvornicanin, Enes. “Accuracy vs AUC in Machine Learning | Baeldung on Computer Science.” *Www.baeldung.com*, 16 Dec. 2021, www.baeldung.com/cs/ml-accuracy-vs-auc#:~:text=Accuracy%20is%20a%20very%20commonly. Accessed 23 May 2022.