Faiza Rafi – ML Final Project

Predicting Autoimmune Disease with Clinical Data

Abstract

Autoimmune conditions are increasingly impacting the general population. Diagnosis of autoimmune conditions is difficult due to extreme variations between patients in multiple aspects and the lack of a current standard classification method. Machine learning algorithms can be trained with clinical data to develop a classifier that can effectively predict autoimmune disease. Of random forest, logistic regression, and support vector classification models trained on data including patient symptoms, vitals, and biomarkers, random forest performs the greatest in predicting disease, closely followed by SVC. Blood panel vitals are among the top features that contribute the greatest to disease prediction.

Introduction

Autoimmune diseases are conditions that are classified by a malfunction of the immune system that causes it to attack an individual's healthy tissues, organs, and cells.[1] Although most conditions are tissue-specific and effects are localized to single organ systems, few are systemic and have an impact across multiple different organs or can be organ-specific. For example, multiple sclerosis severely impacts the nervous system, and its implications of the disease include muscle weakness and reduced cognition.[2] However, a common feature of autoimmune conditions is their activation of inflammatory responses. Additionally, they are often characterized by the presence of antinuclear antibodies (ANAs).[1] The onset of autoimmune conditions is influenced by a variety of factors that have been widely discussed, including genetics and lifestyle. Additional factors that have been more recently studied are epigenetic influences caused by clinical intervention, the environment, and others. For example, the link between gut microbiome and autoimmune conditions has been studied in the light of antibiotic abuse, which is known to severely alter an individual's microbiome.[3] Also, there are studies of repeated ultra-violent light exposure in triggering autoimmune diseases.

Ultimately, it is very difficult to treat autoimmune conditions since there is no foreign body that can be isolated and targeted but rather an element of one's functioning. Treatments currently used involve immunosuppressants, which often have detrimental side effects, or non-steroidal anti-inflammatory drugs like aspirin, which only provide temporary relief.[1] It is equally difficult to diagnose autoimmune diseases as well due to their highly variable nature. Many implications of the diseases can be patient-specific, and two patients with the same disease may not only experience different symptoms but have varying vitals and biomarkers or ANAs.[1,4] For example, systemic lupus erythematosus (SLE) can have musculoskeletal manifestations causing joint pain and skin lesions, or some patients may exclusively experience renal manifestations affecting kidney function.[1,5] Additionally, there are multiple differences in ANAs that are seen in

SLE, like anti-phospholipid antibodies, anti-dsDNA antibodies, or the anti-Smith antibody, which is only present in about 30% of SLE patients.[5] Furthermore, many symptoms of autoimmune conditions often present as those of other conditions, such as fatigue, muscle weakness, and dry eyes, and they may be regarded due to other influences like stress. Because of these reasons, there is still a lack of a comprehensive standard for categorizing and diagnosing autoimmune diseases.[4] However, the incidence of these conditions is on the rise as the National Health and Nutrition Examination Survey has reported a five percent increase in ANA prevalence from the period of the early 1990s to the early 2010s, with the latter ten years alone contributing to a 4.7% percent increase.[4]

Therefore, it is increasingly imperative to develop a method to accurately and efficiently diagnose autoimmune conditions. Not only will it allow for earlier detection of conditions, allowing for earlier intervention, which can minimize health impacts and costs, but it can also help to build a universal database for autoimmune diagnoses.[4] Machine learning methods can be utilized to develop a classifier to predict and diagnose autoimmune conditions. Here, random forest, logistic regression, and support vector classifier models will be applied to the data to determine which classifier performs the best and can most effectively predict the autoimmune disease from the clinical data. Additionally, the models will be used to determine which features are most representative of the data and can be prioritized and potentially used for quick diagnoses. It can be hypothesized the random forest model may perform the best since it generates multiple trees using different subsets of the data allowing it to effectively identify patterns and relationships between the data features while reducing the risk of overfitting. It is likely that the antibody biomarker features will be most important in prediction since they are more directly related to immune response and disease-specific activity, whereas general blood vitals may not provide the same level of specificity or may not be as indicative of a particular disease. However, as discussed earlier, while antibodies profiles can vary within the same disease, they are still likely the most informative and revealing of disease identity.

Data

The data used for this project includes data from 12,500 individuals. There are 116 classes in the target label, which include normal patients (who have no autoimmune disease) or patients with a range of 115 different autoimmune diseases. There were 2,500 normal samples and 10,000 disease samples. There are 79 features in the data, including patient demographics (age and gender), blood panel levels (RBC, WBC, hemoglobin, etc.), antibody markers (anti-IF, anti-RNP, etc.), and symptoms (joint pain, low-grade fever, weight loss, etc.). This data was retrieved from Kaggle.[6]

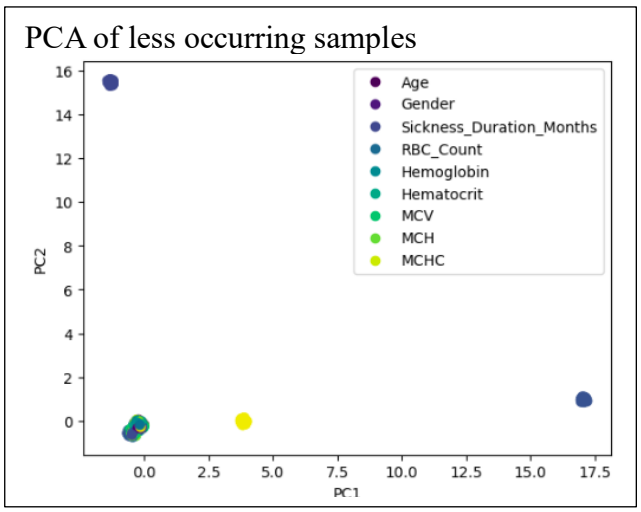
Methods

Data Preprocessing:

Prior to applying the classifiers, the data was processed to clean up the features, remove missing values, etc. First, features of antibody tests with similar names were detected ("Anti-Sm" vs. "Anti_Sm"). Although the values of these "duplicate" features vary, there is no way to confirm that the features are different antibody tests, so they were regarded as duplicates, and the first occurrence of the feature was kept, and all others were removed. There were six features removed. Also, the "Patient ID" feature, which is a number associated with the patient sample, was removed since the integer value can be considered as actual data by the classifiers and skew the results. Next, there were missing values found in seven biomarker features. The data in these features is a binary indicator of the biomarker (1= present, 0 = absent). There were 2,500 null values in these features, so it is assumed that the biomarker test was perhaps not administered in the 2,500 healthy patients. To handle this, the missing values were all converted to zero, indicating the absence of the particular biomarker.

This is a relatively large dataset, and it can be difficult to initially train a classifier to make predictions between 116 different classes in an efficient manner, so the number of classes was reduced. The occurrence of diseases varied from 184 (endometriosis) to 62 (neuromyotonia). To condense the data, top occurring diseases were selected based on a minimum threshold of 90 samples (occurrences), and the remaining diseases were grouped. There were 29 conditions that had an occurrence of at least 90 (this accounts for 25% of the classes), and the total number of samples across the 29 conditions accounted for 44% of the data (5539 samples of 12,500 samples). The remaining less occurring 87 diseases were grouped.

Initial attempts at grouping involved clustering. A PCA was performed with the features of the 87 diseases and determined nine dominant features that contributed the greatest variations in the data. Using the results from the PCA, k-means clustering was performed on the 87 diseases to group them into 10 clusters. However, this resulted in arbitrary grouping.



Below are the results of clusters 0 and 2. Behçet's disease, for example, is seen across clusters (in both 0 and 2) when, ideally, samples of the same disease should congregate in a single cluster. Additionally, some diseases that are clustered into a group do not have much

clinical evidence to support that they are closely related, like Diabetes mellitus and Autoimmune Orchitis (in cluster 2). Ultimately, we can see that this clustering method is not ideal for grouping the diseases. The data contains many binary features, which are likely overlooked by PCA since there is little variation in the values. Also, as discussed earlier, patients with the same disease diagnosis can have vast differences in their vitals and biomarkers influenced by the disease impact itself, clinical intervention, lifestyle, etc.

Cluster 0:

```
clust0= final_data[final_data['Diagnosis'] == 0].index
data.loc[clust0, 'Diagnosis']
```

```
4       Autoimmune polyendocrine syndrome type 2 (APS2)
7                                             Myositis
21                                 Autoimmune retinopathy
22                                     Myasthenia gravis
38                                  Autoimmune oophoritis
                              ...
6938                                   Behçet's disease
6941                        Opsoclonus myoclonus syndrome
6943                                Interstitial cystitis
6947                                 Giant cell arteritis
6951                                           Sarcoidosis
```

Cluster 2:

```
clust2= final_data[final_data['Diagnosis'] == 2].index
data.loc[clust2, 'Diagnosis']
```

```
1                                        Dermatomyositis
3                                Restless legs syndrome
5                                   Autoimmune orchitis
12                                   Addison's disease
14                                    Susac's syndrome
                              ...
6908                                Diabetes mellitus type 1
6927    Undifferentiated connective tissue disease (UCTD)
6931                                   Behçet's disease
6944                                   Behçet's disease
6954                          Discoid lupus erythematosus
```

Therefore, in order to effectively categorize the diseases into groups that can be supported by clinical evidence, the less occurring conditions were categorized by the organ system that they primarily impact. Pathophysiological effects (and some emphasis on symptoms) of the 87 diseases were considered, and they were categorized into nine organ system groups. This resulted in 38 total classes (one for normal patients, 28 disease classes, and 9 grouped disease classes).

After processing the data, there were 12,500 patient samples with 74 features and 38 classes. The data was split into an 80% training set (10,000 samples) and a 20% test set (2,500 samples).

The classifiers (discussed) were applied to this data. However, there is class imbalance in the dataset, the number of samples for each condition varies greatly. Majority classes, like normal, cardiovascular+blood, make up 20% and 11% of the data, respectively, while minority classes, interstitial cystitis, and premature ovarian failure, make up less than 1% of the data. In order to improve the performance of and effectively train the classifiers, this imbalance was handled with two methods on the training set. The RandomUnderSampling[7] method was used to reduce the number of samples in the majority classes. Redundant samples were removed so that the three classes had 1000 samples each. Then, the SMOTE [8] method was used to artificially generate samples in minority classes (35 classes) so that it has the same number of samples as the majority classes.

The final training set to which the classifier was applied has 38,000 samples (previously 10,000), 74 features, and 38 classes. The test set remained unchanged, with 2,500 samples.

Classifiers:

Random forest, logistic regression, and support vector classifier (SVC) models were applied to the training set with GridSearchCV to conduct hyperparametric tuning and identify the best-performing parameter for each model. These models were then fitted to the test set and assessed by ROC-AUC and F1 scores. The best parameters were selected by the highest weighted ROC-AUC score.

Results

With the initial training on the unbalanced data:

Using grid search, random forest, logistic regression, and SVC models were trained with the data setting the 'class_weight' parameter to 'balanced' to allow the classifiers to natively handle the class imbalance in the data.

The random forest model with criterion=gini, max depth = 25, and n_estimators=500 performed the greatest based on the highest ROC-AUC score at 0.783. this test had an F1 score of 0.417.

The logistic regression model with l2 penalty, newton-cg solver, and C = 10, performed the greatest based on the highest ROC-AUC score at 0.788.

The SVC model with gamma scale, rbf kernel, and C = 1, performed the greatest based on the highest ROC-AUC score at 0.786.

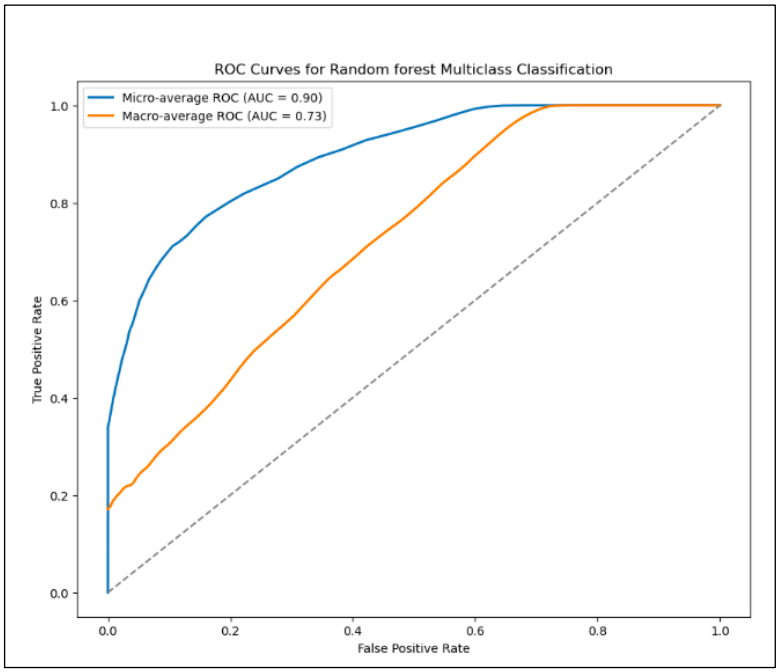
Although accuracy was not used as a primary assessment of model performance, initial tests considered this metric as well to gain a preliminary understanding of model performance. These models' accuracy was around 40%. Given that these scores are from the training set and that the test set can have reduced performance levels, the class imbalance in the training data was resolved, and the models were trained again on balanced data.

With training on the balanced data:

The random forest classifier that performed the best on the training data had the following parameters: gini criterion, max_depth = 50, and n_estimators = 500. The ROC-AUC of this test was 0.989, the F1 was 0.879, and the accuracy was 88%. When this model was fit to the test data, the weighted ROC-AUC score was 0.78. However, the F1 score dropped to 0.4 as a result of the test set still having class imbalance. The ROC curve indicating the micro and micro AUC scores is shown below.

Top-performing random forest model scores on test set

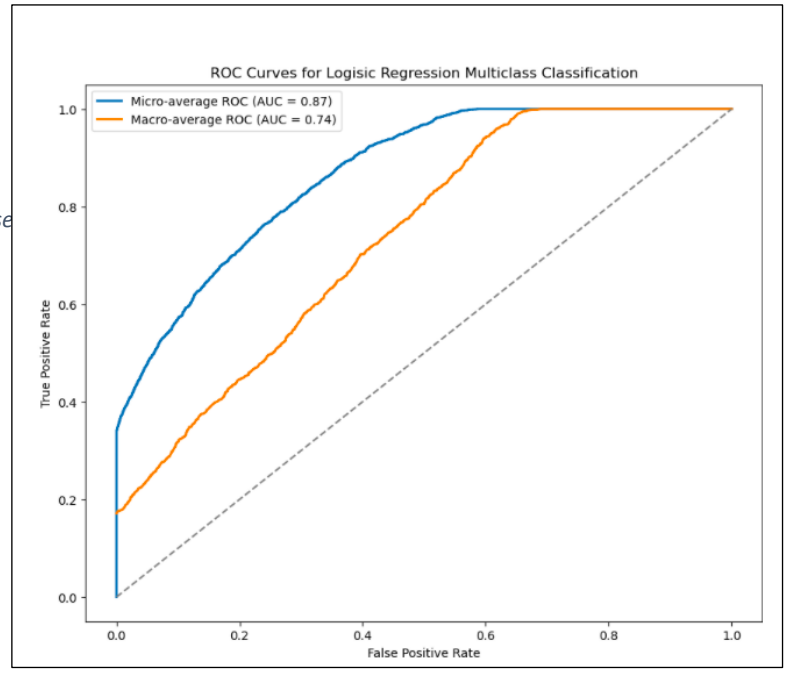
```
RandomForestClassifier(max_depth=50, n_estimators=500, n_jobs=-1,
                       oob_score=True, random_state=1)
parameter: {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 500}
ROC score weighted -- test: 0.7849125580035894
ROC score micro -- test: 0.8963315005405406
ROC score micro -- test: 0.7266843579806137
accuracy: 0.4344
f1: 0.4167052049392143
```



The logistic regression classifier that performed the best on the training data had the following parameters: l2 penalty, lbfgs solver, and C = 0.1. The ROC-AUC of this test was 0.802, the F1 was 0.267, and the accuracy was 27%. When this model was fit to the test data, the weighted ROC-AUC score was 0.79 and the F1 score improved to 0.397. The ROC curve indicating the micro and micro AUC scores is shown below

Top-performing logistic regression model scores on test se

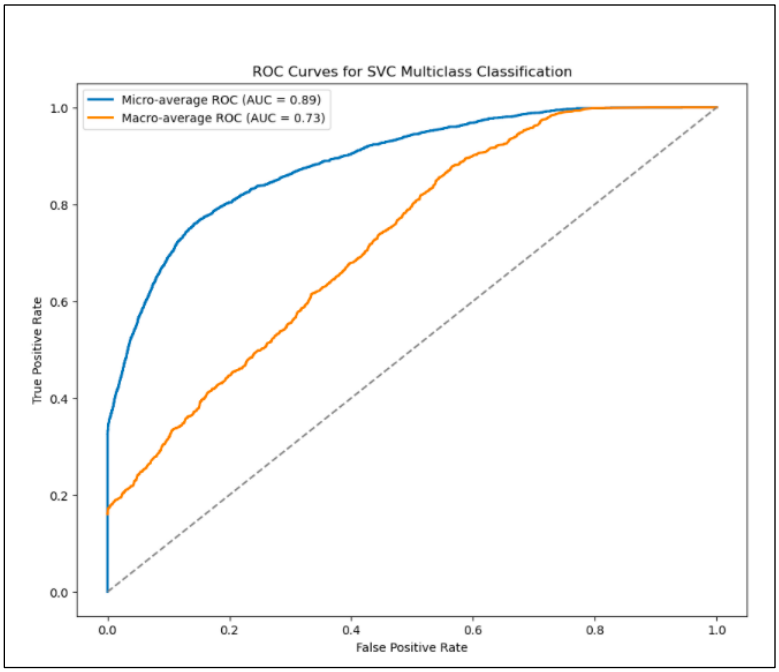
```
LogisticRegression(C=0.1, max_iter=10000, n_jobs=-1)
parameter: {'C': 0.1, 'solver': 'lbfgs'}
ROC score weighted -- test: 0.7900331863441046
ROC score micro -- test: 0.8686485708108109
ROC score micro -- test: 0.7361226664316055
accuracy: 0.3752
f1: 0.39700724346739996
```



The SVC model that performed the best on the training data had the following parameters: gamma scale, rbf kernel, and C =100. The ROC-AUC of this test was 0.987, the F1 was 0.844, and the accuracy was 84%. When this model was fit to the test data, the weighted ROC-AUC score was 0.776 and the F1 score dropped to 0.396. The ROC curve indicating the micro and micro AUC scores is shown below

Top-performing SVC model scores on test set

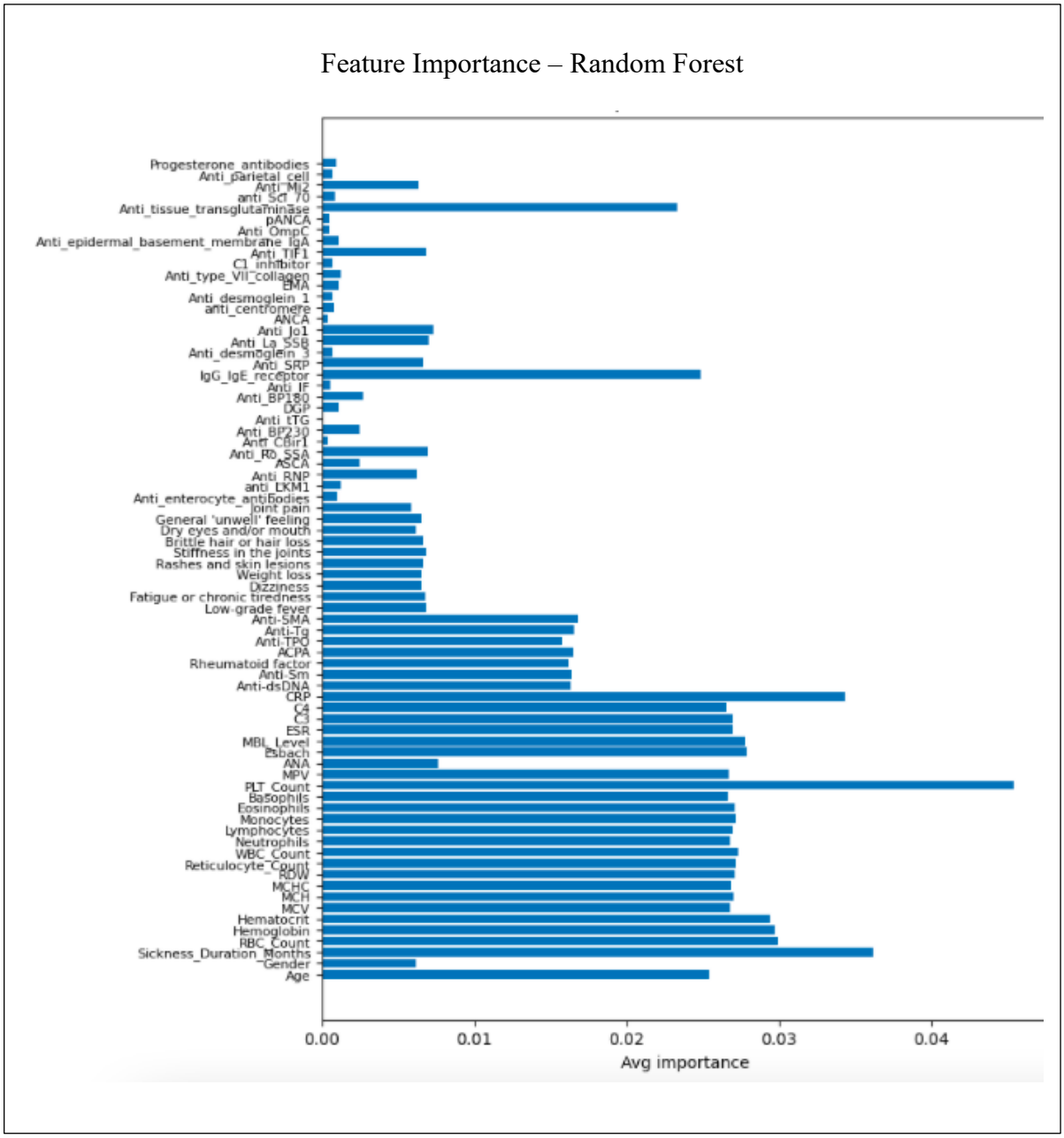
```
SVC(C=100, max_iter=10000, probability=True, random_state=1)
parameter: {'C': 100, 'kernel': 'rbf'}
ROC score weighted-- test: 0.7765571882984801
ROC score micro -- test: 0.8879322162162161
ROC score macro -- test: 0.724784557649255
accuracy: 0.3752
f1: 0.39654572542176475
```



Both the random forest and SVC models perform very similarly in both the training and test sets. The ROC-AUC score of logistic regression was comparable to the other two models, but the F1 score was much lower in the training dataset itself, indicating low performance.

Feature Importance:

Feature contribution was assessed from the random forest classifier. The bar plot of the average importance of each feature is shown below. The top five features that are most significant for predicting disease are PLT count, sickness duration (months), CRP, RBC count, and hemoglobin.



Discussion

The random forest and SVC models yielded very similar results. In the training set, they performed well with ROC-AUC scores at almost 1, F1 scores greater than 0.8, and greater than 80% accuracy. With the test dataset, the classifiers had relatively high weighted ROC-AUC scores of 0.78. However, the F1 scores significantly decreased to 0.4. This is likely due to the fact that the test set had a large class imbalance where the majority class makes up 20% of the data and minority classes account for less than 1%. The lack of minority samples makes the classifiers unable to correctly predict cases of those diseases with few occurrences. Still, both these classifiers can be considered well-performing, with the random forest being a slightly more optimal model and is less prone to overfitting, unlike SVC models. In contrast, the logistic regression model performed at a lower level in comparison to the other two classifiers with the training set; the ROC-AUC score is within an acceptable range, but the F1 is extremely low at 0.3. This immediately establishes that this classifier is not effective for this data since the training set is balanced yet is unable to minimize false positives and negatives.

Additionally, feature importance testing of the random forest classifier revealed that blood panel vital features are more significant in predicting disease than the expected antibody features. This may be again due to the fact discussed earlier that antibodies are not always indicative of autoimmune disease and are often variable within diseases. Also, for the classifier, the antibody features are binary and may not provide as much information to create associations whereas features like hemoglobin, RBC, etc. have a range in the data that may help the classifiers to create nuanced connections between the disease and feature values. Additionally

Above, it was evident that class imbalance significantly impacts the classifiers' performance and, in large datasets with greatly disproportionate classes using the option 'class_weight=balanced' is not sufficient. Thus, the classifiers may be further improved by handling class imbalance in the test set by applying weights to the classes or verifying its performance on another test set that is balanced to avoid generating artificial samples. Also, it may be useful to attempt to cluster classes by blood vital features since there may be significant patterns of those features to the diseases. This can also be used in the future to reduce the features in the data, remove those that contribute noise, and increase the computational power of the classifiers.

References

1. Mu S, Wang W, Liu Q, Ke N, Li H, Sun F, Zhang J, Zhu Z. Autoimmune disease: a view of epigenetics and therapeutic targeting. Front Immunol. 2024 Nov 13;15:1482728. doi: 10.3389/fimmu.2024.1482728. PMID: 39606248; PMCID: PMC11599216.

2. Barkhane Z, Elmadi J, Satish Kumar L, Pugalenthi LS, Ahmad M, Reddy S. Multiple Sclerosis and Autoimmunity: A Veiled Relationship. Cureus. 2022 Apr 19;14(4):e24294. doi: 10.7759/cureus.24294. PMID: 35607574; PMCID: PMC9123335.

3. Khan MF, Wang H. Environmental Exposures and Autoimmune Diseases: Contribution of Gut Microbiome. Front Immunol. 2020 Jan 10;10:3094. doi: 10.3389/fimmu.2019.03094. PMID: 31998327; PMCID: PMC6970196.

4. Justiz Vaillant AA, Goyal A, Varacallo M. Systemic Lupus Erythematosus. [Updated 2023 Aug 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535405/

5. Miller FW. The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention. Curr Opin Immunol. 2023 Feb;80:102266. doi: 10.1016/j.coi.2022.102266. Epub 2022 Nov 26. PMID: 36446151; PMCID: PMC9918670.

6. https://www.kaggle.com/datasets/abdullahragheb/all-autoimmune-disorder-10k/data?select=Final_Balanced_Autoimmune_Disorder_Dataset.csv

7. https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

8. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/