CS 760: Machine Learning Fall 2020

Cell Classification of Severe COVID-19 Patients using Single-Cell RNA-seq Data

AUTHORS: Parth Khatri and Shannon Stiles

Single-cell RNA-sequencing is being used to profile the immune landscape of patients with differing COVID-19 symptoms to try to better understand the different immune cell responses to the viral infection. In this paper a comparison was made of the capability of K-Nearest Neighbors (KNN) and a radial basis function (RBF) Support Vector Machine (SVM), two supervised machine learning methods, to classify cells from Severe COVID-19 and Control donors based on cell gene expression profiles from Single-cell RNA-sequencing data (scRNA-seq). The RBF SVM had the best performance in terms of both accuracy and efficiency. The RBF SVM had 91.8% accuracy, 0.97 sensitivity, and 0.86 specificity, while the KNN classifier had 80.1% accuracy, 0.86 sensitivity, and 0.73 specificity. For both classifiers, all patients in the testing set were correctly identified as Severe COVID-19 or Control based on the accuracy of their cell classifications.

1 Introduction

Background: Coronavirus Disease-2019 (COVID-19) is an enveloped, positive single-strand RNA virus that belongs to the beta-coronavirus genus [1]. COVID-19 is a highly communicable disease that causes severe acute respiratory problems and is transmitted from person-to-person through droplets or direct contact [2]. As of the beginning of December 2020, over 13.6 million Americans have been infected with COVID-19, and there have been 262,293 related deaths [3].The clinical manifestations of COVID-19 are highly variable and include: fever, dry cough, fatigue, loss of taste, and pulmonary problems, like pneumonia [2]. Some patients present mild symptoms and don't require hospitalization, whereas some patients present severe symptoms that lead to hospitalization, intubation, mechanical ventilation, organ support, and death [4]. Patients with severe COVID-19 exhibit an excessive inflammation response [5].

Peripheral Blood Mononuclear Cells (PBMCs) are major cells in the human body's immune response [6]. PBMCs are considered any blood cells with a round nucleus that are isolated from peripheral blood, and which include lymphocytes like T-cells, B-cells, and natural killer cells (NK-cells), monocytes, and dendritic cells [7]. Patients with severe COVID-19 have hyperactivation of monocytes, specifically CD14+ and CD16+ in their peripheral blood, in addition to T-cells becoming pathogenic [5, 8]. Monocytes also undergo morphological and functional changes in patients that have prolonged hospitalization from COVID-19 [9].

Single-cell RNA seq (scRNA-seq) methods have gained substantial traction in molecular biology and genomics over the past 5 years. While still a relatively expensive method compared to bulk RNA-seq or DNA sequencing, scRNA-seq provides high resolution at both the cell-type and potentially the spatial level. It also enables understanding of expression patterns of different cell-types as well as the discovery of rare cell-types. Since this method focuses on RNA, it provides an accurate measure of gene expression in a cell. Single-cell RNA sequencing has been used to investigate the different immune responses in COVID-19 patients and how the disease perturbs the immune cell landscape [10, 11]. There is a decrease in peripheral T-cells and NK-cells in severe COVID-19 patients and an increase in activated T-cells, specifically cytotoxic effector T-cell subsets, whose functional changes can be described through differential gene expression [10, 12, 13].

The gene expression within cells can be used as features in machine learning classifiers to predict if a patient has COVID-19.

Two supervised machine learning methods that can be used for binary classification problems are K-Nearest Neighbors and Support Vector Machines. K-Nearest Neighbors methods are among the simplest machine learning methods for classification tasks. Pre-labeled training data is put into n dimensional space, where n is the number of features considered. Test data is then put into the same space, with labels being assigned based on the k nearest neighbors from the training set. The tunable parameters for KNN methods are k - the number of neighbors evaluated and l - the minimum number of votes (neighbors) required to classify a sample from the test data [14]. Different functions can also be used as metrics to determine the distance between points for evaluation in this method, for example, Euclidean or Manhattan distance. More advanced distance formulations may also be utilized for this purpose [15].

Support Vector Machines (SVMs) are a supervised machine learning method [17]. SVMs can be used to classify both linear and nonlinear data by mapping data to higher dimensional feature space and constructing a hyperplane to separate the data. SVMs minimize empirical classification error, while maximizing the geometric margin between the two classes of data. New samples of data are mapped into the same higher dimensional feature space using kernel methods and classified based on the side of the margin to which they map [18]. The radial basis function (RBF) is a widely used kernel to accommodate a non-linear boundary between classes. It depends on the magnitude of the distance between observations [19]. It is very sensitive to training observations nearby for classifying test observations and observations that are far away have an insignificant role in classification. Gamma is a tuning parameter that stipulates the range of a training observation. A low gamma means the training observation has a wide reach, while a high gamma means the training observation has a narrow reach. Another tuning parameter for SVMs is the cost, C. C determines the tolerance of the model to violations to the margin and hyperplane. A higher C is more tolerant to violations in comparison to a lower C [20].

Related Work: Despite the recency of the COVID-19 crisis, machine learning methods have already been employed on data from COVID-19 patients for the purpose of classifying them. These methods have primarily focused on clinical data that one could obtain in a hospital visit such as image analysis of chest X-rays [21, 22, 23], computed tomography (CT) scans [24], and public health data such as patient demographic, health, geographic, and travel data [25]. There have not been as many attempts to look at sequence data, even less so using scRNA-seq. The cost of scRNA-seq as well as its infrequent clinical usage may account for why its not as widely applicable.

Project Objective: To classify patients as severe COVID-19 or Control based on their cell's scRNA-seq gene expression values comparing two machine learning methods, K-Nearest Neighbors classification and Support Vector Machine classification using a Radial Basis Function (RBF).

2 Data

The COVID-19 PBMC scRNA-seq data was originally from a study published in Cell on September 17, 2020 by Schulte-Schrepping et al. titled *Severe COVID-19 Is Marked by Dysregulated Myeloid Cell Compartment* [26]. The count matrices data was obtained through author permission at: https://www.fastgenomics.org/news/fg-covid-19-cell/. The study consisted of two cohorts of patients in Germany. One cohort was from Berlin and one cohort was from Bonn. The PBMC scRNA-seq data from the Bonn cohort was used in this paper. The scRNA-seq Bonn cohort consisted of a mixture of fresh PBMC, frozen PBMC, and whole blood samples at different sampling dates from 16 Control donors, 8 Mild COVID-19 donors, and 10 Severe COVID-19 donors. Not all of the donors had all three sample types. The scRNA-sequencing platform used for whole transcriptome analyses of the PBMC samples was BD Rhapsody. BD Rhapsody uses a cartridge to capture single-cells and magnetic oligonucleotide barcoded beads for molecular indexing of mRNA transcripts that

are then pooled for library contruction through cDNA amplification [27]. The original count matrices were stored in a Seurat object that consisted of 33,429 features across 139,838 samples prior to preprocessing.

3 Methods

Code was run using R version 4.0.3 and Seurat version 3.2.2 locally on a computer with Windows 64-bit operating system and 16 GB of RAM. The memory limit of Rstudio was extended to approximately 500 GB at the start of the analyses.

3.1 Seurat Preprocessing

Seurat: The Seurat R package was used to perform QC and analysis of the scRNA-seq data prior to creating KNN and SVM classifiers. The scRNA-seq data is encapsulated within the Seurat object [28, 29].

Preprocessing: The fresh and frozen PBMC data for the Bonn cohort were merged, and two patients with only whole blood samples were dropped from the data set. A new metadata variable was made to define if the sample was from a COVID-19 or Control donor. Cells with less than 200 unique feature counts and over 2,500 were filtered out along with cells that had greater than 15% mitochrondrial counts. Seurat data processing was performed with built-in Seurat functions. The data was scaled globally with Seurat's "LogNormalize" method. Feature expression measurements were normalized by total expression, multiplied by a scale factor, and log transformed. 2000 features, i.e. genes, that were highly variable between cells were subset to be used downstream in the Principal Component Analysis (PCA). A linear transformation was performed by shifting the expression of each gene so that the mean expression was 0 and scaling the expression of each gene so that the variance across cells was 1. PCA was performed to linearly reduce the dimensions of the data using the variable genes. The dimensionality of the data set was determined to be 17 through an Elbow Plot visualization, which can be seen in Figure 3 in the Appendix[28, 29, 30]. The total number of genes, i.e., features in the cell count matrices were 33,419. A workflow was originally designed to accommodate all 33,419 features for training the KNN and SVM classifier, but due to computational inefficiency and memory limits an alternative workflow was designed using a differential gene expression analysis to select a subset of the features to use.

Differential Gene Expression Analysis A K-Nearest Neighbors graph was constructed using Euclidean distance in the PCA space with 17 dimensions, then the cells were clustered using modularity optimization techniques within Seurat's *FindClusters* function with a defined resolution of 0.5. A Uniform Manifold Approximation Projection (UMAP) was run to perform non-linear dimensional reduction. The UMAPs were visualized by mild, severe, and control donor groups separately and overlaid over each other. The differentially expressed genes between cells that came from Severe COVID-19 donors and Control donors were found using the *FindMarkers* function in Seurat [28, 29, 31]. 1,122 genes were found with significantly different expressions between the two groups. The list of genes with p-values can be found in a supplementary csv file in the data folder.

3.2 Defining Training and Testing Data Sets

Cell count matrices were extracted for the 8 Severe COVID-19 donors within the Seurat data object and 8 randomly selected Control donors. Each donor was assigned a corresponding label that was added to the dataframe in a new column along with donor ID. Severe COVID-19 donors were assigned 1 and Control donors were assigned -1. The cell count dataframes for all donors were bound together in a master dataframe. The dataframe was separated into Control and COVID-19 dataframes and the donors were split randomly

using a seed of 21 to the training and testing data sets. The training and testing data sets each contained 4 Severe COVID-19 donors and 4 Control donors. The training data set contained 19,492 cell samples and the testing data set contained 19,451 cell samples. Demographic information about the donors assigned to each data set can be seen in Table 4 in the Appendix. The training and testing datasets were refined to only the differentially expressed genes. The number of genes, i.e., features was reduced from 33,419 to 1,122 prior to model training.

3.3 Cell Classification

K-Nearest Neighbors: The training data, testing data, and training labels were passed into R's *knn* function from the class package with the k parameter set to $\sqrt{n}$, where n is the number of samples in the training data, to predict the classifications of the testing data based on the common convention for choosing k[33]. Euclidean distance was used to compute distance to the nearest neighbors and majority vote was used to determine the classification. For tie breaks, all candidates were included in the vote [32]. The system.time function was used to evaluate runtime efficiency. The prediction results were evaluated by creating a confusion matrix with the correct labels to visualize the number of true positives, false positives, true negatives, and false negatives and calculate the accuracy, sensitivity, and specificity. A receiver operating characteristic (ROC) curve was plotted using the ROCit package to visualize the Sensitivity (True Positive Rate) versus 1 - Specificity (False Positive Rate) rate and the area under the curve (AUC) was calculated [34].

Radial Basis Function Support Vector Machine: The training data and labels were passed into R's *svm* function from the e1071 package with the *type* parameter set to "C-Classification," the *kernel* parameter set to "radial", the *gamma* parameter set to 0.001, and the *cost* parameter set to 0.1 to train a radial basis function SVM classifier. An optimization experiment was performed to determine the optimal C and gamma parameters from a C range of 0.01, 0.1, 1, and 10 and a gamma range of 0.000001, 0.001, 1, and 10. The optimization results can be seen in Table 5 in the Appendix. The *predict* function was used to predict the labels of the testing data passed in using the classification model trained by *svm* [35]. The system.time function was used to evaluate runtime efficiency. The prediction results were evaluated by creating a confusion matrix with the correct labels to visualize the number of true positives, false positives, true negatives, and false negatives and calculate the accuracy, sensitivity, and specificity. A ROC curve was plotted with the ROCit package to visualize the true positive rate versus false positive rate and AUC was calculated [34].

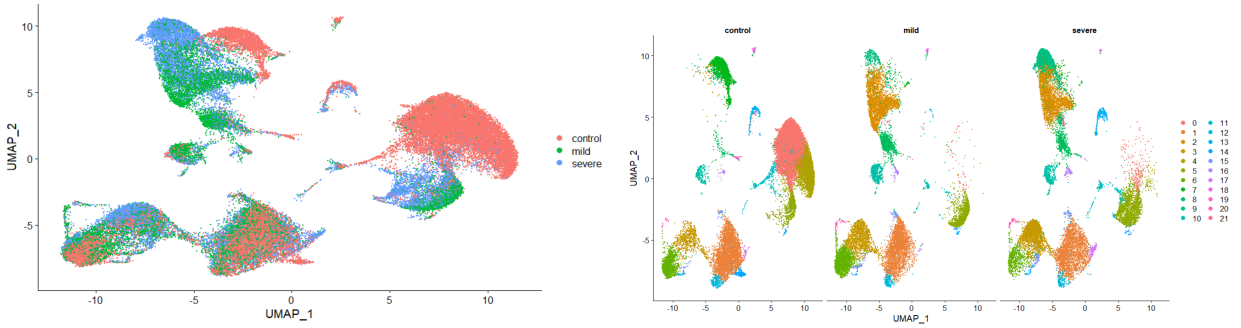
3.4 Patient Classification

The donor id, condition label, and the prediction for each cell was reimposed on the testing set. For each donor in the testing set, the testing set was filtered to only include the cells of the patient. The correct label and the predictions were summarized in a table, the number of correctly and incorrectly classified cells were pulled out, and the accuracy was calculated. Donors labeled as *Severe COVID-19* with more than 50% of their cells correctly classified were predicted to have *Severe COVID-19*. Donors labeled as *Control* with more than 50% of their cells correctly classified were predicted to be a *Control*.

4 Results

4.1 Computational Limits to Whole scRNA-seq Dataset

A training and testing set each consisting of 8 COVID-19 donors (either mild or severe) and 8 Control donors was originally defined to look at all COVID-19 samples regardless of severity in comparison to



(a) Overlaid Cell Distributions

(b) Separate Cell Distributions

Figure 1: UMAPs for Mild, Severe, and Control Cells. There are distinct cell types that are from Control donors and distinct cell types from COVID-19 donors. The cells from mild COVID-19 donors are slightly more disperse than the cells form the severe COVID-19 donors. Cells in group two are primarily from COVID-19 donors and cells in group 21 are mostly from Control donors.

Control samples. The memory within Rstudio limited the ability to perform this classification task. A memory allocation error occurred that could not be remedied by extending the memory.limit in Rstudio. A subset of the 8 severe COVID-19 donors and 8 randomly sampled Controls were separated into training and testing sets with all 33,419 gene features and an attempt to train the KNN and RBF SVM classifiers was made. The attempt was aborted when it took longer than 22 hours for the KNN classifier to train and over 6 hours for the RBF SVM model to train. The testing and training data sets were refined to only the 1,122 differentially expressed genes as features to improve computational efficiency and feasibility.

4.2 Differentially Expressed Genes

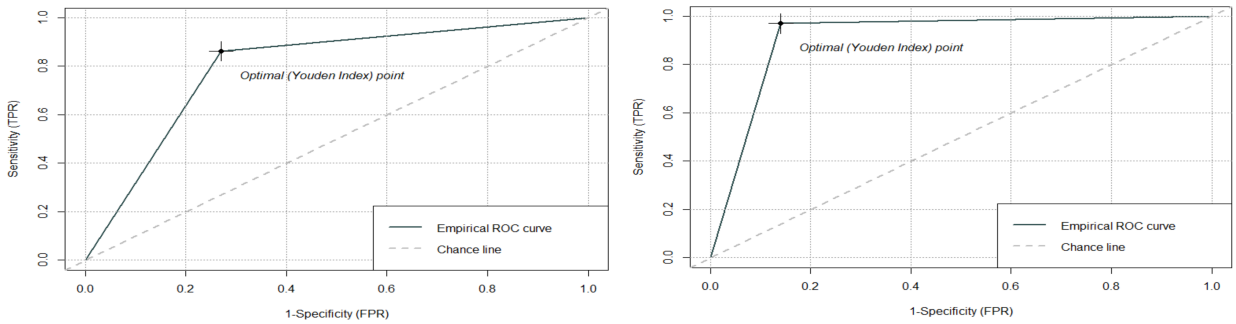
There were 1,122 genes that had significantly different expression between the cells from severe COVID-19 donors and the cells from Control donors. The whole list of differentially expressed genes can be seen in the *Severe Control DE genes* csv in the project repository's data folder. The overlaid cell distributions show unique cell types from the Control donors separate from the severe and mild COVID-19 donors as seen in Figure 1(a). There is some slight overlap between the cells from the mild and severe COVID-19 donors, but the cells from mild donors are more disperse. Specifically, the severe and control groups are of interest based on the donors in the data sets. In Figure 1 (b), it can be seen that the cells from group two are almost exclusively from COVID-19 donors, while the cells from group 21 are almost exclusively from Control donors.

4.3 Cell Classification

The KNN classifier took approximately 106 minutes to run. Its prediction accuracy was approximately 80%, while it had a sensitivity of around 0.86 and a specificity around 0.73. The RBF SVM was more efficient and had a much shorter runtime. The RBF SVM classifier took approximately 17 minutes to train the model and predict the labels of the test set. 8,940 support vectors were used in the model. 4,477 support vectors corresponded to cells labelled as Severe COVID-19 and 4,463 support vectors corresponded to cells labelled as Control. Its prediction accuracy, sensitivity, and specificity were better than the KNN classifier. The RBF SVM's prediction accuracy was approximately 92%, its sensitivity was approximately 0.97, and its specificity was approximately 0.86. The results of both classifiers are summarized in Table 1.

Classifier	Time (min)	Accuracy (%)	Sensitivity	Specificity	TP	FP	TN	FN
KNN	106	80.09314	0.8624344	0.7298698	9034	2449	6617	1441
RBF SVM	17	91.84791	0.9693556	0.8596956	10154	1272	7794	321

Table 1: Results of Machine Learning Classifier Comparison of Cell Classification



(a) KNN ROC Curve

(b) RBF SVM ROC Curve

Figure 2: Receiver Operating Characteristic (ROC) curves for the KNN and RBF SVM classification of cells from Severe COVID-19 and Control donors. The optimal point in the SVM ROC curve shows better specificity in comparison to the KNN ROC curve. The sensitivity at the optimal points for both ROC curves are similar. The area under the curve (AUC) value for the KNN classifier is 0.7961521 and the AUC value for the SVM classifier is 0.9145256. The SVM model is a better classifier than the KNN model.

The ROC curves for each classifier also show the higher sensitivity and specificity in the RBF SVM classifier in comparison to the KNN classifier as seen in Figure 2. The optimal point for the RBF SVM shows that 1-specificity is less than 0.15 in comparison to the KNN ROC optimal point which has 1-specificity greater than 0.2 and that sensitivity is close to 1 in comparison to the KNN sensitivity which is around 0.85. The AUC value for the RBF SVM was approximately 0.915, which was better than the AUC value of approximately 0.796 for the KNN ROC curve.

4.4 Patient Classification

All 4 Severe COVID-19 donors and 4 Control donors were correctly classified by both the KNN and RBF SVM cell classification predictions. The criteria defined that if a Severe COVID-19 donor had over 50% of the their cells classified correctly as COVID-19 by the respective model, they were classified as Severe COVID-19, and it defined that if a Control donor had over 50% of their cells classified as Control by the respective model, they were classified as Control. The patient classification results for the KNN classifier are summarized in Table 2 and the results for the RBF SVM classifier are summarized in Table 3.

The RBF SVM classifier had better accuracy overall in comparison to the KNN classifier for all donors. The classification accuracy for the RBF SVM was less variable in comparison to the KNN. The range of cell accuracy values for all of the donors was approximately 30% for the KNN classification results and approximately 22% for the RBF SVM classification results.

Donor	Number of Cells	Condition	Accuracy (%)	Correct	Incorrect	Prediction
BN-12	2434	Severe COVID-19	95.48069	2324	110	Severe COVID-19
BN-15	3051	Severe COVID-19	92.55982	2824	227	Severe COVID-19
BN-16	2290	Severe COVID-19	67.11790	1537	753	Severe COVID-19
BN-17	2700	Severe COVID-19	87.00000	2349	351	Severe COVID-19
BN-22	3166	Control	71.57296	2266	900	Control
BN-27	542	Control	76.56827	415	127	Control
BN-29	2066	Control	82.86544	1712	354	Control
BN-31	3292	Control	67.55772	2224	1068	Control

Table 2: Results of KNN Patient Classification

Donor	Number of Cells	Condition	Accuracy (%)	Correct	Incorrect	Prediction
BN-12	2434	Severe COVID-19	93.79622	2283	151	Severe COVID-19
BN-15	3051	Severe COVID-19	99.01672	3021	30	Severe COVID-19
BN-16	2290	Severe COVID-19	94.71616	2169	121	Severe COVID-19
BN-17	2700	Severe COVID-19	99.29630	2681	19	Severe COVID-19
BN-22	3166	Control	77.00569	2438	728	Control
BN-27	542	Control	86.90037	471	71	Control
BN-29	2066	Control	91.81994	1897	169	Control
BN-31	3292	Control	90.76549	2988	304	Control

Table 3: Results of RBF SVM Patient Classification

5 Discussion

Classifier Comparison: The RBF SVM was capable of finding a non-linear separation between the 19,492 training cell samples using a subset of points, specifically, 4,477 support vectors for severe COVID-19 cells and 4,463 support vectors for Control cells. The RBF SVM was more efficient than the KNN classifier, since it classifies test data points based on the region where it is located with respect to the hyperplane boundary [39]. The KNN classifier calculates the distance between each testing data point and all of the data points in the training set, which is computationally inefficient for large data sets [36]. The KNN had a longer runtime than the RBF SVM. Both the KNN and RBF SVM are sensitive to outliers, although the SVM is slightly more robust to outliers, since the gamma and C parameters can be tuned to limit the range of observations used to classify a testing data point and the tolerance for misclassified points [19]. The KNN and RBF SVM classifiers could be improved through more parameter tuning. 16 different combinations of gamma and C parameters were explored for the RBF SVM classifier in an optimization experiment prior to the final RBF SVM parameter choice as seen in Table 5 in the Appendix. The number of tuning combinations explored was limited by local runtime efficiency. The RBF SVM could possibly be improved by expanding the range of the gamma and C parameters used. The KNN classifier could also be improved by exploring different values of k to optimize the accuracy of the classifier.

Feature Selection: The large feature space of the original data set that included expression data from over 33,000 genes caused memory allocation problems and infeasible runtimes. After refining the workflow with a feature selection approach, the classifiers ran efficiently and yielded good results. A differential gene expression analysis was used to select genes that had significantly different expression between Severe COVID-19 and Control donor cells and reduced the feature space from 33,419 genes to 1,122 genes. Interestingly, many of the genes identified in immune cell profile analyses of COVID-19 patients in the literature were

identified through the differential gene expression analysis and selected as features. Many genes associated with cytotoxicity that were identified as highly expressed in COVID-19 patients by Zhang et al. in 2020 had differential gene expression between Severe COVID-19 and Control donor cells in the data set used in this paper and were selected as features [12]. These genes include: NKG7, GZMA, GZMB, GZMH, and GNLY. Zhu et al. found IL6R upregulated in PBMCs, which causes an increase in proinflammatory cytokines during pathogensis. The also found upregulated of ISG15, AFI44, FI44L, and RSAD2 which are interferon-stimulated genes [11]. All of these genes were selected as features in this paper based on their differential expression profiles. This corroboration of differentially expressed genes with the literature provides support for the feature selection approach applied in this paper. The application of the differential gene expression analysis on all severe COVID-19 and Control cells did introduce some bias into the classifiers, since the feature selection was based on all samples including testing samples and not just the training samples. Essentially, the current approach allowed 'peaking' into gene signatures associated with the test set. To minimize bias, the data set should be split into training and testing prior to the differential gene analysis, and the features selected should only come from genes that have significantly different expression between the COVID-19 and Control donors in the training set.

Future Directions: The KNN and RBF SVM classifiers in this paper could be enhanced by scRNA-seq specific additions to the data set. The first would be the inclusion of more donor samples, like cohort 1 from Schulte-Schrepping et al. Cohort 1 samples used the 10x genomics scRNA-seq method, rather than the BD Rhapsody one. The integration of the two data sets is possible through the Harmony package in R which groups cells by type rather than conditions specific to technologies or experimental conditions [37, 38]. The genes measured across different data sets could also differ, so discussions on how to integrate data sources while also finding common differentially expressed genes for feature selection are important. The capability to predict COVID-19 status at the individual cell level could be improved by utilizing cell type as a variable for prediction. Different cell types have large differences in expressed genes, which would also be exhibited in cells from patients infected with COVID-19. Different cells that drive the immune response have different ratios in patients with and without COVID-19, specifically neutrophil-to-lymphocyte ratio[10]. Integrating different cell types like neutrophils and tissue types from the same cohort into the PBMC data would be an interesting next step. Extending the data set to different COVID-19 patient samples taken from a wider range of geographic areas may also indicate different strains of COVID-19 that exhibit unique differential expression patterns, which would be valuable for the classifier.

Lastly, there were three classes of donors in the data set: Mild COVID-19, Severe COVID-19, and Control donors. Future work on the classifier could find differences in gene expression in each of the combinations of types of patients as an attempt to make a multiclass classifier. The severity of the symptoms could be predicted as a 3-class problem as listed above or a 7-class problem that would use the WHO score of a patient [4]. More patient data as well as a more robust cross validation process like Leave One Out cross validation or 5-fold cross validation would improve and help confirm the findings presented in this paper.

Conclusion: This project provided a good use case for comparing two supervised machine learning methods learned in CS760 to perhaps the world's most pressing current problem. Patients with and without severe COVID-19 were able to be correctly classified based on the classification predictions of their cells from both KNN and RBF SVM classifiers using gene expression data from differentially expressed genes. The RBF SVM classifier is recommended to classify whether cells came from a donor with or without severe COVID-19 based on its accuracy and efficiency. The RBF SVM had an accuracy around 91.8%, a sensitivity around 0.97, and a specificity around 0.86, while the KNN classifier had an accuracy around 80.1%, a sensitivity of around 0.86 and a specificity around 0.73.

All work for the project can be found here: Project Repository and Large Data File Storage

References

[1] Y. C. Wu, C. S. Chen, and Y. J. Chan, "The outbreak of COVID-19: An overview," Journal of the Chinese Medical Association, vol. 83, no. 3. Wolters Kluwer Health, pp. 217–220, 2020, doi: 10.1097/JCMA.0000000000000270.

[2] Y. Shi et al., "An overview of COVID-19," Journal of Zhejiang University: Science B, vol. 21, no. 5. Zhejiang University Press, pp. 343–360, May 01, 2020, doi: 10.1631/jzus.B2000083.

[3] "US Historical Data — The COVID Tracking Project." https://covidtracking.com/data/national (accessed Dec. 05, 2020).

[4] World Health Organization, "WHO RD Blueprint novel Coronavirus COVID-19 Therapeutic Trial Synopsis," World Health Organization, no. February 18, 2020, Geneva, Switzerland, pp. 1–9, 2020, [Online]. Available: http://www.moh.gov.sa/en/CoronaNew/PressReleases/Pages/default.aspx.

[5] M. Merad and J. C. Martin, "Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages," Nature Reviews Immunology, vol. 20, no. 6, pp. 355–362, 2020, doi: 10.1038/s41577-020-0331-4.

[6] J. Pourahmad and A. Salimi, "Isolated human peripheral blood mononuclear cell (PBMC), a cost effective tool for predicting immunosuppressive effects of drugs and Xenobiotics," Iranian Journal of Pharmaceutical Research, vol. 14, no. 4. Iranian Journal of Pharmaceutical Research, pp. 679–980, Sep. 01, 2015, doi: 10.22037/ijpr.2015.1790.

[7] C. Kleiveland, "Peripheral blood mononuclear cells," in The Impact of Food Bioactives on Health: In Vitro and Ex Vivo Models, Springer International Publishing, 2015, pp. 161–167.

[8] Y. Zhou et al., "Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients," National Science Review, vol. 7, no. 6, pp. 998–1002, 2020, doi: 10.1093/nsr/nwaa041.

[9] D. Zhang et al., "COVID-19 infection induces readily detectable morphologic and inflammation-related phenotypic changes in peripheral blood monocytes," Journal of Leukocyte Biology, p. JLB.4HI0720-470R, Oct. 2020, doi: 10.1002/JLB.4HI0720-470R.

[10] G. Xu et al., "The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing," Cell Discovery, vol. 6, no. 1, 2020, doi: 10.1038/s41421-020-00225-2.

[11] L. Zhu et al., "Single-Cell Sequencing of Peripheral Mononuclear Cells Reveals Distinct Immune Response Landscapes of COVID-19 and Influenza Patients," Immunity, vol. 53, no. 3, pp. 685-696.e3, 2020, doi: 10.1016/j.immuni.2020.07.009.

[12] J. Y. Zhang et al., "Single-cell landscape of immunological responses in patients with COVID-19," Nature Immunology, vol. 21, no. 9, pp. 1107–1118, 2020, doi: 10.1038/s41590-020-0762-x.

[13] W. Wen et al., "Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing," Cell Discovery, vol. 6, no. 1, 2020, doi: 10.1038/s41421-020-0168-9.

[14] B.D. Ripley, "Pattern Recognition and Neural Networks," Cambridge: Cambridge University Press, 1996.

[15] W. Venables and B. Ripley, "Modern Applied Statistics with S Fourth edition," 4th ed. Springer, 2002.

[16] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," vol. 12, no. 8, pp. 33–39, 2014, [Online]. Available: http://arxiv.org/abs/1409.0919.

[17] C.M. Bishop, "Kernel Methods," *Pattern Recognition and Machine Learning*, 6ht ed. New York, USA: Springer, 2006, ch. 6, pp. 291-301.

[18] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," 2012. Accessed: Nov. 11, 2020. [Online].

[19] C.M. Bishop, "Kernel Methods," Pattern Recognition and Machine Learning, 6th ed. New York, USA: Springer, 2006, ch. 6, pp. 292, 299-301.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Support Vector Machines," An Introduction to Statistical Learning, by New York, USA: Springer, 2017, ch. 9, pp. 347-353.

[21] A. Zargari Khuzani, M. Heidari, and A. Shariati, "COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images," medRxiv: the preprint server for health sciences, no. Ml, 2020, doi: 10.1101/2020.05.09.20096560.

[22] A. T. Sahlol, D. Yousri, A. A. Ewees, M. A. A. Al-qaness, R. Damasevicius, and M. A. Elaziz, "COVID-19 image classification using deep features and fractional-order marine predators algorithm," Scientific Reports, vol. 10, no. 1, pp. 1–15, 2020, doi: 10.1038/s41598-020-71294-2.

[23] R. Zhang et al., "Diagnosis of COVID-19 Pneumonia Using Chest Radiography: Value of Artificial Intelligence," vol. 78, no. September, pp. 1–15, 2020.

[24] T. D. Pham, "A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks," Scientific Reports, vol. 10, no. 1, pp. 1–8, 2020, doi: 10.1038/s41598-020-74164-z.

[25] C. Iwendi et al., "COVID-19 patient health prediction using boosted random forest algorithm," Frontiers in Public Health, vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[26] J. Schulte-Schrepping et al., "Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment," Cell, vol. 182, no. 6, pp. 1419-1440.e23, 2020, doi: 10.1016/j.cell.2020.08.001.

[27] E. Y. Shum, E. M. Walczak, C. Chang, and H. Christina Fan, "Quantitation of mRNA Transcripts and Proteins Using the BD RhapsodyTM Single-Cell Analysis System," in Advances in Experimental Medicine and Biology, vol. 1129, Springer New York LLC, 2019, pp. 63–79.

[28] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," Nature Biotechnology, vol. 36, no. 5, pp. 411–420, Jun. 2018, doi: 10.1038/nbt.4096.

[29] T. Stuart et al., "Comprehensive Integration of Single-Cell Data," Cell, vol. 177, no. 7, pp. 1888-1902.e21, 2019, doi: 10.1016/j.cell.2019.05.031.

[30] R. Satija, "Seurat - Guided Clustering Tutorial." https://satijalab.org/seurat/v3.1/pbmc3ktutorial.html (accessed Dec. 07, 2020).

[31] R. Satija, "Tutorial: Integrating stimulated vs. control PBMC datasets to learn cell-type specific responses." https://satijalab.org/seurat/v3.2/immunealignment.html (accessed Dec. 07, 2020).

[32] A. Brian, W. Venables, and M. B. Ripley, "Package ' class ,'" 2020. [Online]. Available: https://cran.r-project.org/web/packages/class/class.pdf.

[33] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," vol. 12, no. 8, pp. 33–39, 2014, [Online]. Available: http://arxiv.org/abs/1409.0919.

[34] "CRAN - Package ROCit." https://cran.r-project.org/web/packages/ROCit/index.html (accessed Dec. 11, 2020).

[35] K. Hornik, A. Weingessel, F. Leisch, and D. Meyer, "Package 'e1071,'" 2020. [Online]. Available: https://cran.r-project.org/web/packages/e1071/e1071.pdf.

[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Assessing Model Accuracy," An Introduction to Statistical Learning, by New York, USA: Springer, 2017, ch. 2, pp. 39.

[37] I. Korsunsky et al., "Fast, sensitive and accurate integration of single-cell data with Harmony," Nature Methods, vol. 16, no. 12, pp. 1289–1296, Dec. 2019, doi: 10.1038/s41592-019-0619-0.

[38] "immunogenomics/harmony: Fast, sensitive and accurate integration of single-cell data with Harmony." https://github.com/immunogenomics/harmony (accessed Dec. 11, 2020).

[39] D. Pimentel-Alarcon, "Topic 11: Support Vector Machines," pp. 1–13, 2020, [Online]. Available: https://danielpimentel.github.io/teaching/CS760/lectures/CS76011SVMs.pdf.

6 Appendix

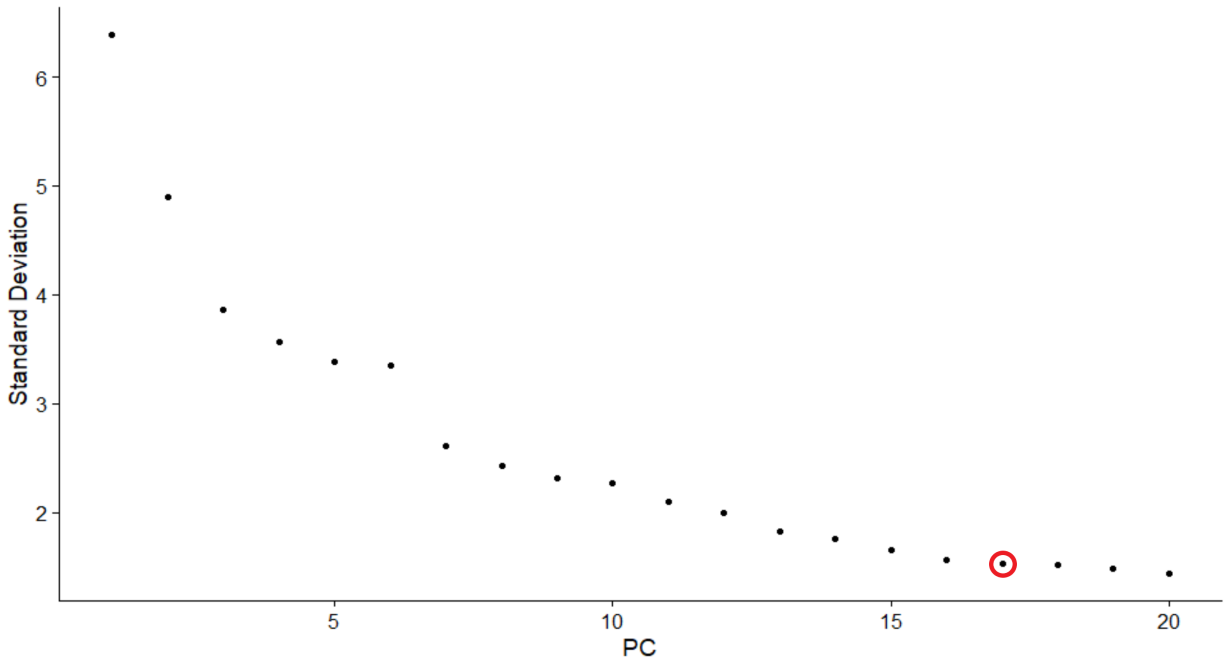


Figure 3: The dimensions of the scRNA-seq data set was evaluated with an Elbow Plot after PCA in Seurat. The dimensions were determined to be 17.

Donor	Dataset	Condition	WHO Classification	Number of Cells	Sex	Age Range (years)
BN-11	Training	Severe COVID-19	7	2276	Male	51-55
BN-12	Testing	Severe COVID-19	3	2434	Female	61-65
BN-14	Training	Severe COVID-19	7	3221	Female	81-85
BN-15	Testing	Severe COVID-19	7	3051	Male	71-75
BN-16	Testing	Severe COVID-19	7	2290	Female	81-85
BN-17	Testing	Severe COVID-19	7	2700	Male	81-85
BN-18	Training	Severe COVID-19	7	1027	Female	76-80
BN-19	Training	Severe COVID-19	5/7	2621	Female	61-65
BN-22	Testing	Control	N/A	3166	Male	46-50
BN-23	Training	Control	N/A	2145	Male	51-55
BN-24	Training	Control	N/A	2897	Male	51-55
BN-26	Training	Control	N/A	3516	Female	46-50
BN-27	Testing	Control	N/A	542	Female	61-65
BN-28	Training	Control	N/A	1789	Female	56-60
BN-29	Testing	Control	N/A	2066	Female	66-70
BN-31	Testing	Control	N/A	3292	Male	61-65

Table 4: Processed Donor Data Set and Demographic Information. 8 Donors has Severe COVID-19 Symptoms and 8 Donors were Controls without COVID-19. The Severe COVID-19 donors for the most part had a WHO Classification of 7, which means that they are on Ventilation and organ support. A score of 5 means that the patient is on Non-invasive ventilation or high-flow oxygen. A score of 3 means the patient has been hospitalized [4].

C	Gamma	Accuracy (%)	Sensitivity	Specificity
0.01	0.000001	46.39476	0	1
0.01	0.001	77.22225	0.8873508	0.6392014
0.01	1	46.39476	0	1
0.01	10	46.39476	0	1
0.1	0.000001	46.39476	0	1
0.1	**0.001**	**91.84791**	**0.9693556**	**0.8596956**
0.1	1	46.39476	0	1
0.1	10	46.39476	0	1
1	0.000001	75.05245	0.8480191	0.6378778
1	0.001	89.27895	0.8624344	0.9278623
1	1	46.39476	0	1
1	10	46.39476	0	1
10	0.000001	88.69045	0.9811933	0.7779616
10	0.001	88.62392	0.8481146	0.9302890
10	1	46.39476	0	1
10	10	46.39476	0	1
1	$\frac{1}{Numberof Features}$	*89.22266*	*0.8613842*	*0.9278623*

Table 5: RBF SVM Hyperparameter Optimization. 16 hyperparameter combinations with C ranging form 0.01, 0.1, 1, and 10 and Gamma ranging from 0.000001, 0.001, 1, and 10 were explored for the RBF SVM classifier. The default hyperparameters for the *svm* function (in italics in the last row) were also tested. The hyperparameter combination that yields the best accuracy is C=0.1 and Gamma=0.001 (in bold above). The accuracy was 92% in comparison to the default accuracy of 89%. The sensitivity was better in comparison too. Maximizing true positives (high sensitivity) in diagnosing COVID-19 is important so that people who are positive are identified and do not spread the disease undetected.