Machine Learning Approaches to Predict COVID-19 Infection Severity Using scRNA-seq Data

Charles A. Ellis, Donghyun (Ethan) Kim, Martina Lapera, Max Liu, Abhinav Sattiraju, Carolina Urzay
Georgia Institute of Technology

***Abstract*--Single cell RNA sequencing (scRNA-seq) is a useful technique for gene expression examination at the single-cell level to investigate differentiating factors between cells. In the context of the ongoing COVID-19 pandemic, scRNA-seq could reveal molecular indicators of varying severity of COVID-19 infection. Predicting COVID-19 infection using scRNA-seq data could prove highly valuable for the timely determination of treatment as well as the identification of genes correlated with infection. This paper investigates multiple machine learning methods – Support Vector Machines, Random Forests, and Multilayer Perceptron Models – applied to scBALF-COVID-19, a scRNA-seq dataset provided by the IEEE Healthcare Summit which classifies patients as having either Normal, Mild, or Severe COVID-19 infection status. The models are optimized for a Mean Receiver Operating Characteristic score with an emphasis on explainability, with insights from feature importance analysis. Random Forests were found to yield the best Precision, Recall, and F1-scores greater than 85%. Using permutation importance across multiple Random Forest classification folds, we found 17 genes that were reproducibly associated with COVID-19 severity. A preliminary literature survey supports the validity of some of the genes identified by our approach as being directly related to COVID-19 while also supporting the possibility that we have identified some genes that were not previously known to be associated with COVID-19. The positive results from model training and evaluation corroborate the potential impact that scRNA-seq has for assessing the severity of widespread diseases in the future and could enable improved prediction of COVID-19 severity.**

***Index Terms*--scRNA-seq, COVID-19, Random Forest, Explainable Artificial Intelligence**

I. Introduction

The COVID-19 pandemic has had a tremendous impact over the past year and a half, with over 4.5 million deaths and over 200 million diagnosed cases [1]. It has had drastic consequences on not only the physical health of those affected, but also on those who have lost access to healthcare services because of the saturation of hospitals and other medical centers. Furthermore, the impacts of the pandemic have extended far beyond health-related issues; the pandemic has also had grave effects on people's leisure and work environment, politics in both global and national scales, and the economy of all countries.

An early diagnosis of the virus has been one of the major priorities since the beginning of the pandemic, both to ensure the health of those afflicted, as well as to prevent new cases from emerging. This is also an important step in working towards the discovery of a treatment. Understanding the molecular signatures of the disease provides key information about how it functions and consequently brings researchers closer to a cure. It can give insights on disease mechanism, risk and progression, cell's biology and response to drugs or physiological toxicity, to mention a few examples [2][3].

One way to achieve this is through single-cell RNA sequencing (scRNA-seq), which consists of the detection of the messenger RNA molecules inside a cell [4]. Messenger RNA molecules store sequences of nucleotides that will later be translated into proteins inside the nucleus by the ribosomes. Proteins are molecules involved in cell function, structure and regulation [5]. Thus, if their synthesis was altered by a virus (SARS-CoV-2), the affected individual would suffer from the associated illness (COVID-19).

The mechanism of infection by SARS-CoV-2 is initiated with the binding of the viral spike proteins (S) or surface glycoproteins to the extracellular receptor ACE2 (present on the outer membrane of various cell types in the human body). This interaction will ultimately lead to fusion at the cellular membrane which will enable the invasion of SARS-CoV-2 [6]. Once inside, the virus will make use of the cell's transcription mechanisms in order to synthesize mRNA molecules. These will then be translated into viral proteins that will alter the normal functioning of patients' cells.

The rapid advancement of artificial intelligence and growth of explainability research in recent years has had a strong impact upon bioinformatics analyses. Furthermore, explainable artificial intelligence methods have been applied to scRNA-seq analysis in a number of contexts, such as identifying gene expression and better depicting relationships among scRNA-seq data in an Alzheimer's disease study [7]. As such, scRNA-seq paired with explainability approaches has great potential to be used for identifying genetic markers of COVID-19 symptom severity.

This study examines the viability of using explainable machine learning methods to predict COVID-19 symptom severity. Additionally, and through the use of an explainability approach, key genetic signatures associated with COVID-19 severity were identified.

II. Methods

In this section, we describe our dataset, classification approaches, and explainability approach.

A. Description of Dataset

In this study, we used the scBALF-COVID-19 dataset. The dataset was collected via single cell RNA-seq (scRNA-seq). It contains 1,999 genes and 23,189 samples. Based upon their COVID-19 symptom severity, each sample had an associated label of Normal, Mild, or Severe, with each class composing 52%, 14%, and 34% of the dataset, respectively. Feature-wise z-scoring was performed prior to the training of the classifier.

B. Classification Approach

This study considers three different classification models: random forest (RF), support vector machine (SVM) and multilayer perceptron neural networks (MLP).

When training our models, we used five fold cross-validation. The distribution of each class was maintained when being separated into training, validation, and test sets. Training, validation, and test sets each made up 60%, 20% , and 20%, respectively. The sample splits were standardized across classifiers such that each fold of each model was trained and assessed upon the same samples.

We implemented an SVM using *Scikit-learn*. From Grid Search Cross Validation, we found that a regularization parameter value of 1.5 and a radial basis function kernel produced the best results. Default values were selected for all other parameters. During each fold, we randomly upsampled the minority classes of our training data before training our SVM model.

We implemented an RF using *Scikit-learn.* From Grid Search Cross Validation, we selected a total of 10 trees. Class weights were set to balanced, and default values were selected for all other parameters.

We implemented a MLP in *Keras*. Our architecture consisted of 2 dense layers with 28 nodes, each immediately followed by ReLu activation functions and 40% dropout layers. After the final dropout layer, we had an output layer with 3 nodes and a softmax activation. We used Kaiming He initialization for the model weights and zeros for bias initialization. Additionally, we used the Adam optimizer with a learning rate of 0.0005 and categorical cross entropy loss. We trained for 50 epochs. To account for the effects of the class imbalance, we weighted the loss function of the model.

After training each of the models, we calculated the mean and standard deviation of their sensitivity, precision, and F1-score across all folds. We then output a receiver operating characteristic curve for the model with the top F1-score.

C. Explainability Approach

Permutation feature importance was first introduced for the use with random forests and was later expanded for use with other machine learning classifiers [8]. We used permutation feature importance to gain insight into the gene signatures associated with each fold of the top performing classifier. Permutation importance involves iteratively permuting one feature at a time and examining the effect of the permutation upon the performance of the classifier for that data. We applied the permutation feature importance technique for all folds and identified the top 100 genes for each fold. We then identified the genes that were common across folds.

III. Results and Discussion

In this section, we present and discuss our classification results and the genes that we were able to identify.

A. Classification Results

TABLE I

Performance metrics for class: normal

	Precision (%)	Recall (%)	F1 (%)
RF	93.64 ± 0.90	99.53 ± 0.25	96.50 ± 0.53
SVM	94.93 ± 0.36	92.42 ± 0.60	93.66 ± 0.38
MLP	89.21 ± 0.33	94.16 ± 1.17	91.61 ± 0.51

TABLE II

Performance metrics for class: mild

	Precision (%)	Recall (%)	F1 (%)
RF	97.71 ± 1.01	75.50 ± 3.95	85.16 ± 2.67
SVM	78.96 ± 1.57	83.37 ± 1.90	81.10 ± 1.65
MLP	82.06 ± 2.21	62.46 ± 1.73	70.90 ± 1.49

TABLE III

Performance metrics for class: severe

	Precision (%)	Recall (%)	F1 (%)
RF	99.97 ± 0.06	99.90 ± 0.13	99.94 ± 0.06
SVM	94.87 ± 0.28	96.46 ± 0.60	95.66 ± 0.27
MLP	92.00 ± 0.97	93.38 ± 1.11	92.68 ± 0.57

Tables I-III show the mean and standard deviation of the precision, recall, and F1-score of each class (Normal, Mild, Severe) and each of the classifiers across folds. Sensitivity or Recall measures the ratio of true positives to the sum of true positives and false negatives. Precision measures the ratio of true positives to the sum of true positives and false positives.

F1-score is the harmonic mean of both sensitivity and precision.

The lowest F1 metrics for the RF, SVM and MLP correspond to the Mild COVID-19 Infection Severity class. In fact, overall, for the Mild COVID-19 class, the precision, recall, and F1-score were not as high as for the other classes for all models. This is attributed to the imbalance in the classes, with the Mild class being the most underrepresented. Additionally, it is also feasible that the nucleotide sequences present in the Mild cases are not as distinct from one another as Normal and Severe cases. In contrast to the Mild class, the metrics for all models were the highest for the Severe COVID-19 class.

As indicated by Tables I-III, except for the Precision metric for the Normal COVID-19 class and the Recall metric for the Mild COVID-19 class, the Random Forest Classifier yielded the best results. Indeed, the ROC curve shows optimal results. Thus, we proceeded with the Random Forests for the explainability approach.

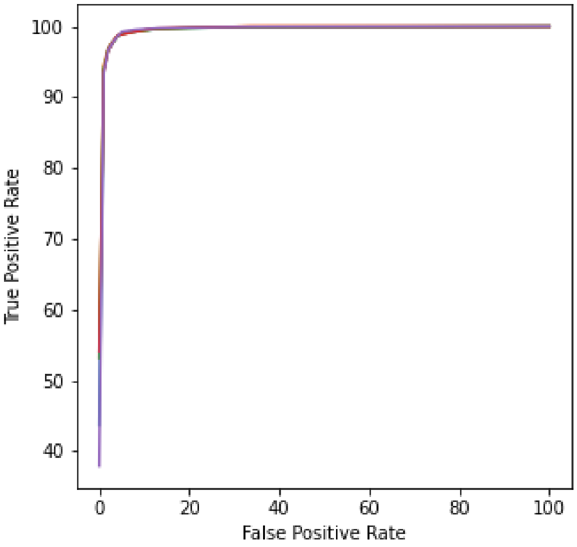

Fig. 1. ROC curve for the Random Forest Classifier.

B. *Identification of genes with highest feature importance (Explainability)*

When we examined the genes that were in the top 5% of genes across all folds, we identified 17 features. Fig. 2 shows the mean permutation importance for each gene. Of the 17 genes, the genes C1orf56, SNHG25, PHKG1, and AC114760.2 had the highest mean feature importance values and can thus be considered molecular markers (molecular signatures) of COVID-19 severity.

SNHG25 gene is a lnRNA (long non-coding RNA). These are molecules proved to be related to the modulation of host-virus interactions in BAL cells [9] (Bronchoalveolar Lavage cells), whose fluid patterns are representative of inflammatory cell profiles in lung tissues [10].

PHKG1 is a gene that codes for a protein of the kinase family: Phosphorylase Kinase Catalytic Subunit Gamma 1. The latter is essential in the pathway that leads to the activation of c-AMP-Dependent PKA [11]. In fact, a recent study claims that *"cAMP-dependent activation of PKA attenuates respiratory syncytial virus-induced human airway epithelial barrier disruption"* [12].

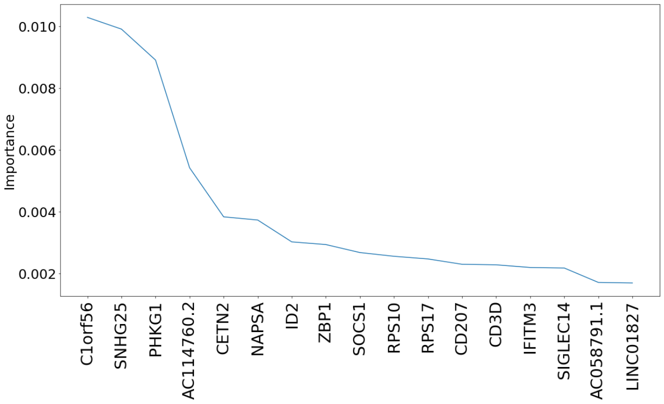

Fig. 2. Feature (gene) importance.

C1orf56 is a proto-oncogene which is repressed through methylation of its promoter by DNMT3B (DNA methyltransferase 3B). [13] No direct relationship between C1orf56 and SARS-CoV-2 was found, however, DNMT3B gene was classified in the green list of genes directly related to COVID-19. [14]

Thus, overall, biologically relevant results related to COVID-19 development in patients were obtained from the genes with the top three highest feature importance.

C. *Limitations and Next Steps*

While our overall methods were sound, they do have some limitations resulting from computation times. For our Random Forest, we obtained very high performance. However, to maximize the reliability of the explainability results, it would have been better to allow each tree to pick from all features rather than randomly assigning a small subset of features for building each tree. Additionally, due to the high data dimensionality of the data, we were unable to run permutation feature importance for a large number of repeats. In the future, using a larger number of cross-validation folds and a larger number of repeats could be helpful. Furthermore, clustering analysis could help with marker identification as well. Finally, one possible cause of lower performance for the Mild COVID-19 Infection Severity class is that it contains less values than the other 2 classes (imbalance in data). If the model were trained on a more balanced dataset, this issue might be resolved.

V. Conclusion

The COVID-19 pandemic has had a monumental effect on the global population, leaving millions deceased and a social, economic, and political crisis behind in its wake. It has highlighted the importance of understanding organisms on a molecular scale to aid in the effort of preventing widespread diseases in the future. Single-cell RNA sequencing provides a useful tool for analyzing genetic signatures, which can provide for the examination of relationships between cells and indicate varying levels of COVID-19 severity. This study takes advantage of explainable machine learning techniques to find differences between the genes expressed in 3 levels of COVID-19 infection severity: Normal, Mild, and Severe. Of the Support Vector Machine, Random Forest, and Multilayer Perceptron Model, the Random Forest was found to yield the best performance metrics on the data, with an F1-score of 99.9% for the Severe COVID-19 class. A mean ROC curve was also developed from the Random Forest results over the 3 classes. From permuting feature importance on the Random Forests, 17 genes were identified to be associated with COVID-19 severity. Of those 17, 4 are known to be molecular markers of COVID-19 severity. Since we drew our metrics and feature importance scores from five-fold cross-validation, our results can be easily reproduced. This could have positive implications on further research into analyzing scRNA-seq data to prevent future widespread, deadly diseases.

ACKNOWLEDGMENTS

The authors gratefully thank the IEEE Healthcare Summit for the opportunity to participate in the COVID-19 Data Hackathon.

References

[1] "Coronavirus (COVID-19)," *Google News*. [Online]. Available: https://news.google.com/covid19/map?hl=en-US&mid=%2Fm%2F02j71&gl=US&ceid=US%3Aen. [Accessed: 03-Oct-2021].

[2] J. Sung, Y. Wang, S. Chandrasekaran, D. M. Witten, and N. D. Price, "Molecular signatures from OMICS DATA: From chaos to consensus," *Biotechnology journal*, Aug-2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418428/. [Accessed: 03-Oct-2021].

[3] R. Nilsson, J. Björkegren, and J. Tegnér, "On reliable discovery of molecular signatures," *BMC Bioinformatics*, 29-Jan-2009. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-38. [Accessed: 03-Oct-2021].

[4] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for Biomedical Research and Clinical Applications," *Genome Medicine*, 18-Aug-2017. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4. [Accessed: 03-Oct-2021].

[5] "What are proteins and what do they do?: Medlineplus genetics," *MedlinePlus*, 26-Mar-2021. [Online]. Available: https://medlineplus.gov/genetics/understanding/howgeneswork/protein/. [Accessed: 03-Oct-2021].

[6] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, and V. Thiel, "Coronavirus biology and replication: implications for SARS-CoV-2," *Nature*, Oct. 2020.

[7] Wang, J., Ma, A., Chang, Y. *et al.* scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat Commun* 12, 1882 (2021). https://doi.org/10.1038/s41467-021-22197-x

[8] Breiman, Leo."Random Forests." Machine Learning 45 (1). Springer: 5-32 (2001)

[9] M. Heron, J. C. Grutters, K. M. ten Dam-Molenkamp, D. Hijdra, A. van Heugten-Roeling, A. M. Claessen, H. J. Ruven, J. M. van den Bosch, and H. van Velzen-Blad, "Bronchoalveolar lavage cell pattern from healthy human lung," *Clinical & Experimental Immunology*, vol. 167, no. 3, pp. 523–531, 2012.

[10] M. Heron, J. C. Grutters, K. M. ten Dam-Molenkamp, D. Hijdra, A. van Heugten-Roeling, A. M. E. Claessen, H. J. T. Ruven, J. M. M. van den Bosch, and H. van Velzen-Blad, "Bronchoalveolar lavage cell pattern from healthy human lung," *Clinical and experimental immunology*, Mar-2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3374285/. [Accessed: 03-Oct-2021].

[11] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PHKG1. [Accessed: 03-Oct-2021].

[12] F. Rezaee, T. J. Harford, D. T. Linfield, G. Altawallbeh, R. J. Midura, A. I. Ivanov, and G. Piedimonte, "CAMP-dependent activation of protein kinase a attenuates respiratory syncytial virus-induced human airway epithelial barrier disruption," PLOS ONE, vol. 12, no. 7, 2017.

[13] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=C1orf56. [Accessed: 03-Oct-2021].

[14] *Gene: DNMT3B (COVID-19 research)*. [Online]. Available: https://panelapp.genomicsengland.co.uk/panels/111/gene/DNMT3B/. [Accessed: 03-Oct-2021].