Comparing Differential Gene Expression Related to Covid-19 Status and Severity

Abby's Mentor Group

—

Nathan Lopez
Juan Molina
Divya Nitin
Jack Perkins
Dharsan Selvakumar
Daniel Yang

Introduction:

As of April 2022, the Covid-19 pandemic has infected more than half a billion people and killed nearly seven million more worldwide (World Health Organization, 2022). The Covid-19 infection stems from the SARS-CoV-2 and often results in respiratory symptoms such as coughs, shortness of breath, and pneumonia (Guan et. al 2020). Since the start of the pandemic, studies have identified various biological markers such as transcripts, proteins, and metabolites associated with Covid-19 related respiratory infections (Ellinghaus et al., 2020), aiding doctors in identifying, treating, and performing triage on these patients.

One of the key challenges facing doctors regarding the pandemic is determining the cause of the respiratory infection. Since Covid-19 symptoms mimic many common respiratory symptoms, it is difficult to distinguish a potential Covid-19 infection from non-Covid-19 respiratory infection with similar symptoms. In addition, Covid-19's effects can range from mild to very severe, and other than age and pre existing medical conditions, it has been difficult to identify reliable predictors for a patient's outcome.

Furthermore, predicting the severity of a case, regardless of the cause, would benefit doctors as they could begin treatment sooner on patients that need it more. Our work classified patients suffering from respiratory infections into Covid or non-Covid, and determined if the patient will need ICU care through identifying differential gene expression. By identifying the genes associated with Covid and ICU status, doctors will be able to more efficiently determine the best course of care needed for individual patients.

Methods:

We used data that was publicly available and provided by a previous study (Overmyer, 2021) consisting of 128 blood samples from patients who came to the emergency room with COVID-19-like symptoms. Of those who tested positive for COVID-19, 51 people were admitted into the ICU and 51 were not. Of those who tested negative for COVID-19, 16 were admitted into the ICU and 11 were not.

	Covid	Non-Covid	Total
ICU	50	16	67
Non-ICU	50	10	61
Total	100	26	126

Table 1. Data table showing number of covid and non-covid samples in our study. They were split into ICU and Non-ICU.

We first went through several steps of preprocessing of the data. First, the TPM (Transcripts Per Million) normalized data was obtained from the Overmyer study. In order to prepare the data for statistical analysis, we log-transformed the data. In addition, we filtered the lowest counts data using quantile filtering, getting rid of rows with more than 80% of values

being zero. Using this quantile filtering, we reduced the dataset from 19,472 genes to 16,169 genes.

PCA (Principal Component Analysis) was conducted to visualize sources of variation in the data. We wanted the genes that would most clearly show the difference between the 4 different groups. For this reason, the 25th percentile of highest variation genes were selected, resulting in 4042 genes used. A PCA chart then showed the visualized data based on COVID status and another PCA chart to differentiate based on ICU status. Lastly, 2 variables were combined, creating a 4 way PCA chart to differentiate between the 4 possibilities.

Next we conducted a 4-way anova test to compare the four groups specified in table 1. We then created a heatmap with these groups and analyzed the data clusters formed by the dendrograms. We also determined which genes were differentially expressed when it came to Covid and ICU status and used an alpha value of 0.01 for our hypothesis. We used post-hoc testing to compare the Covid and ICU conditions separately and narrow down our results. We analyzed the log2 fold change across Covid and Non-Covid patients as well as ICU and non-ICU patients to find differentially expressed genes specific to these groups. Lastly, the differentially expressed genes were inputted into Enrichr to find the overexpressed gene pathways.

Additionally, the filtered and log-transformed genes were used to create various classification models to be able to predict the patient's Covid-19 status and severity level. The models that we created for our data were a KNN model, a Decision Tree model, and an SVM model. For the KNN model, k was set to 10, following the heuristic of $k=\sqrt{n}$, and for the SVM model, we utilized a linear kernel function for classification. After creating each model, the precision and recall for each model was compared to determine the best model for classification. Each model was tested using 6-fold cross validation on the data set to mitigate the effects of overfitting. To ensure all three models used the same randomization, we set the random state to the same value. Lastly, we combined the differentially expressed genes identified through the 4-way anova tests with the best classification model. By filtering out genes that did not contribute towards classification, the model became more resilient to noise. Lastly, to recreate the Overmyer study more closely, we retrained the best classification model on only Covid status.

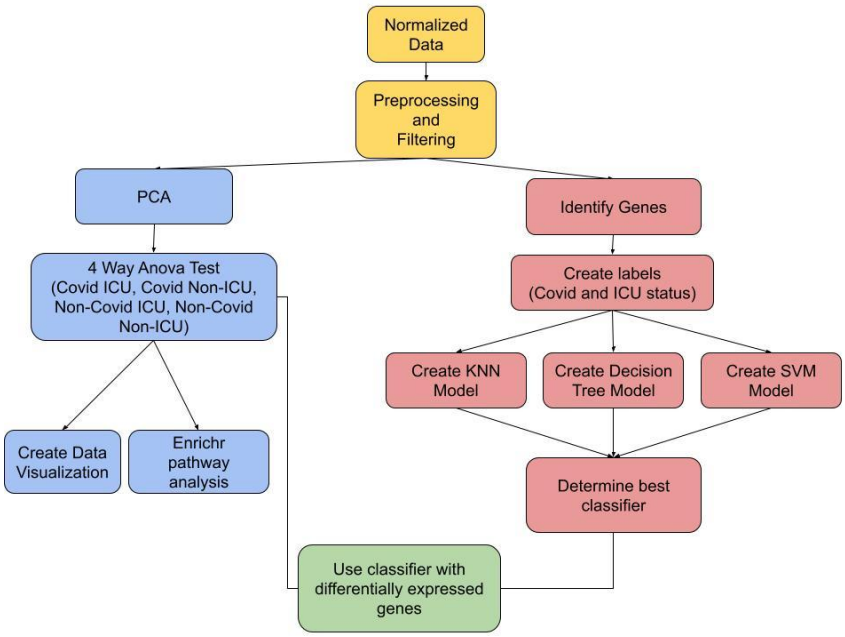

Figure 1. Analysis workflow. The workflow consisted of two main portions: classification (steps shown in red) and differential expression analysis (shown in blue). In the final step (shown in green), the genes identified through differential expression analysis were used as features for the best classification algorithm.

Results:

Principal Component Analysis:

In addition to the classification models that we created, we used Principal Component Analysis. The figures below show the results from the PCA visualizations that we created from our data points with the top 25% most variation.

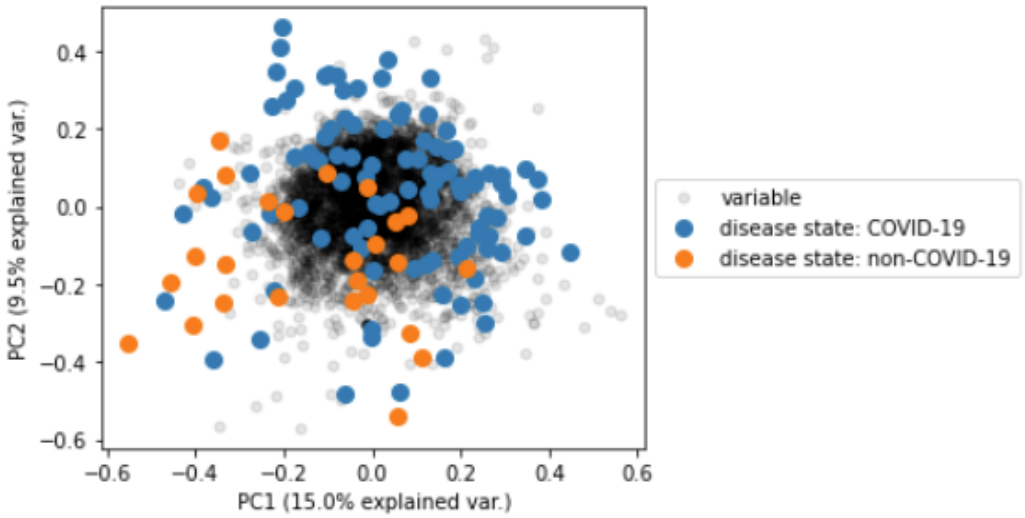

Figure 2. Comparison of the Covid-19 state. Principal Component Analysis of the individual's disease status when looking at the top 25% highest varying genes. The blue points represent patients who tested positive for Covid-19 and the orange points represent patients who tested negative for Covid-19.

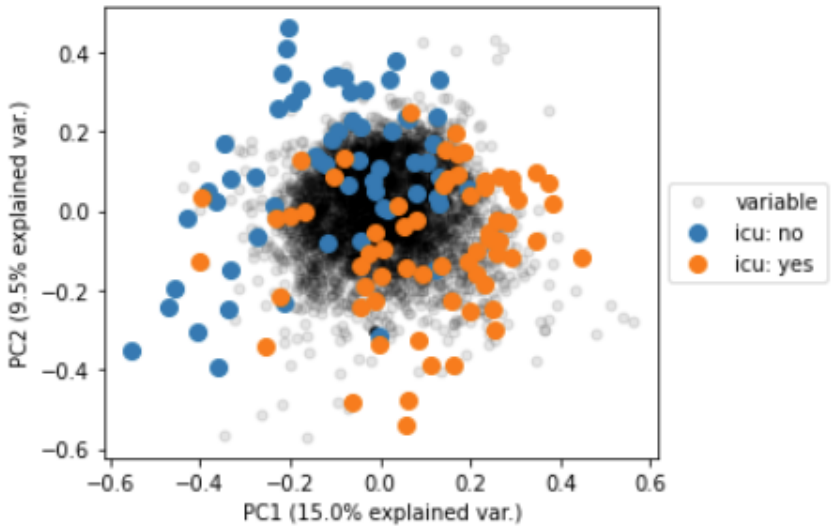

Figure 3. Comparison of the ICU state. Principal Component Analysis of the individual's ICU status when looking at the top 25% highest varying genes. The blue points represent patients who were not placed in the ICU and the orange points represent patients who were placed in the ICU.

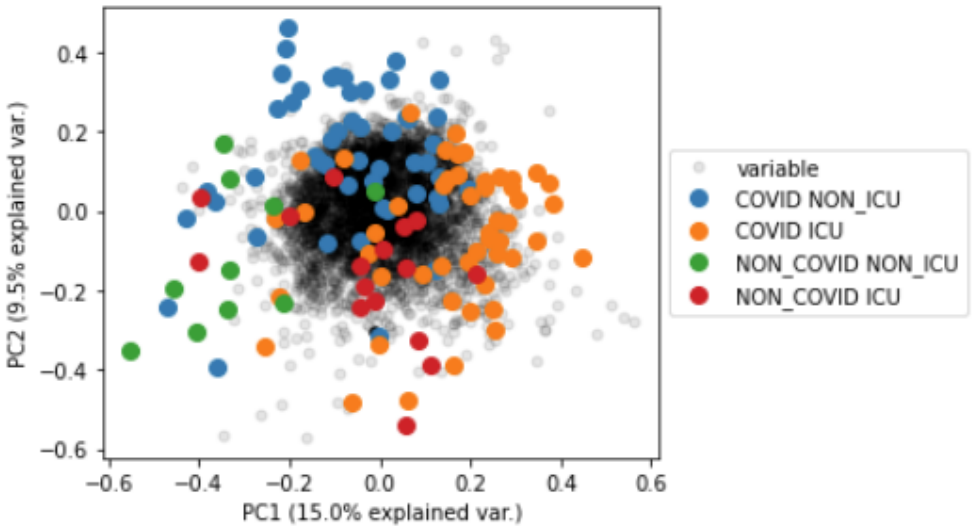

Figure 4. Comparison of the Covid-19 and ICU states. Principal Component Analysis of the individual's Covid-19 and ICU status when looking at the top 25% highest varying genes. The blue points represent patients who tested positive for Covid-19 and were not placed in the ICU.

The orange points represent patients who tested positive for Covid-19 and were placed in the ICU. The green points represent patients who tested negative for Covid-19 and were not placed in the ICU. The red points represent patients who tested negative for Covid-19 and were placed in the ICU.

Differential Gene Expression:

A 4 Way Anova test was created in order to find the differentially expressed genes. With a p-value cutoff of <0.01 and an absolute log2 fold change cutoff of >1.5, 335 differentially expressed genes were identified. The heatmap below was then created to differentiate between COVID and Non-COVID samples.

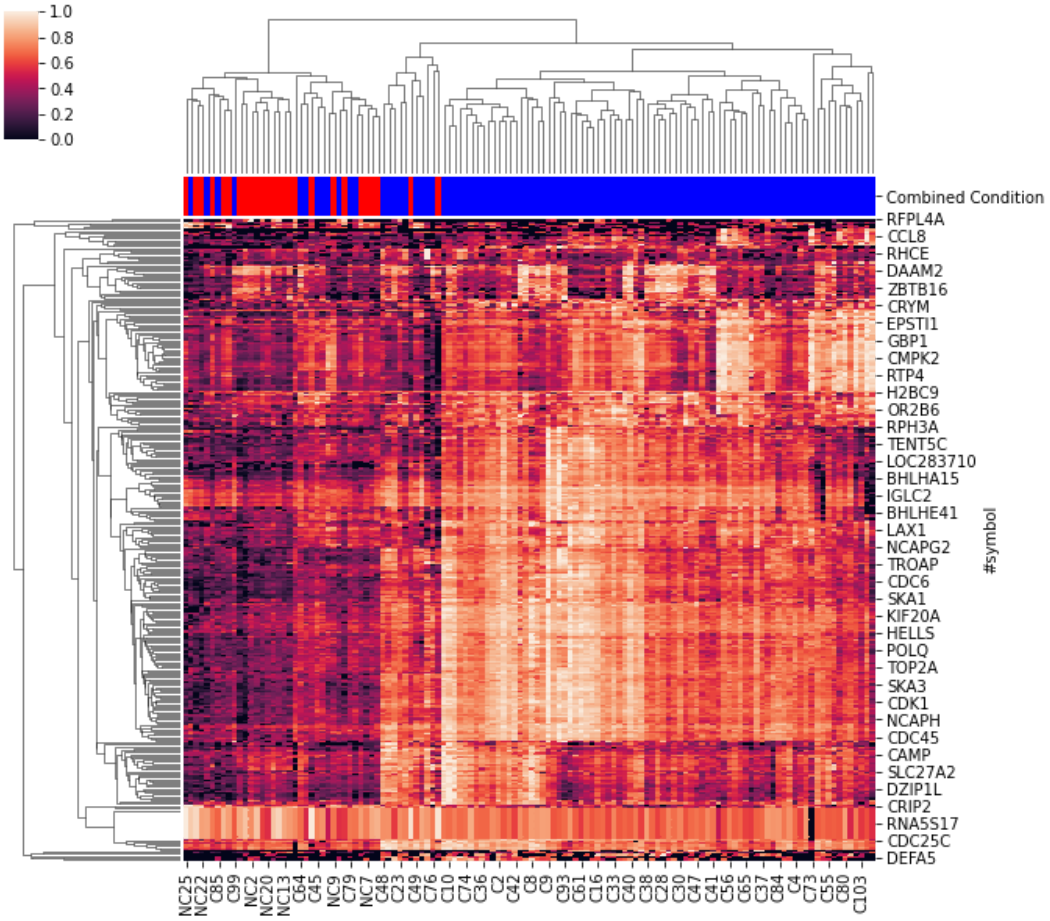

Figure 5: Heatmap with DEGs: Heatmap visualizing the 335 differentially expressed genes, and how the genes cluster according to their gene expression. Darker shades indicate lower expression, while lighter shades indicate higher expression. The color bar at the top refers to the condition of the specific sample: red is Non-COVID and blue is COVID.The dendrograms show the hierarchical clustering of the gene expression levels in the heatmap. The horizontal dendrograms show clustering of specific genes while the vertical dendrograms show clustering

of specific samples.

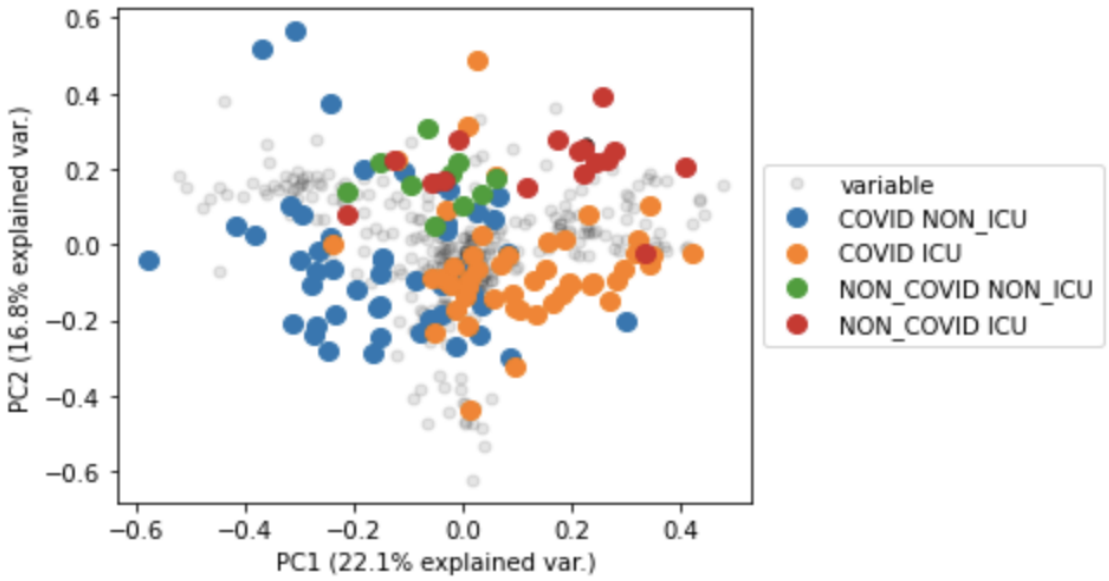

Figure 6. Comparison of the Covid-19 and ICU states. Principal Component Analysis of the individual's Covid-19 and ICU status after identifying differential genes. The blue points represent patients who tested positive for Covid-19 and were not placed in the ICU. The orange points represent patients who tested positive for Covid-19 and were placed in the ICU. The green points represent patients who tested negative for Covid-19 and were not placed in the ICU. The red points represent patients who tested negative for Covid-19 and were placed in the ICU.

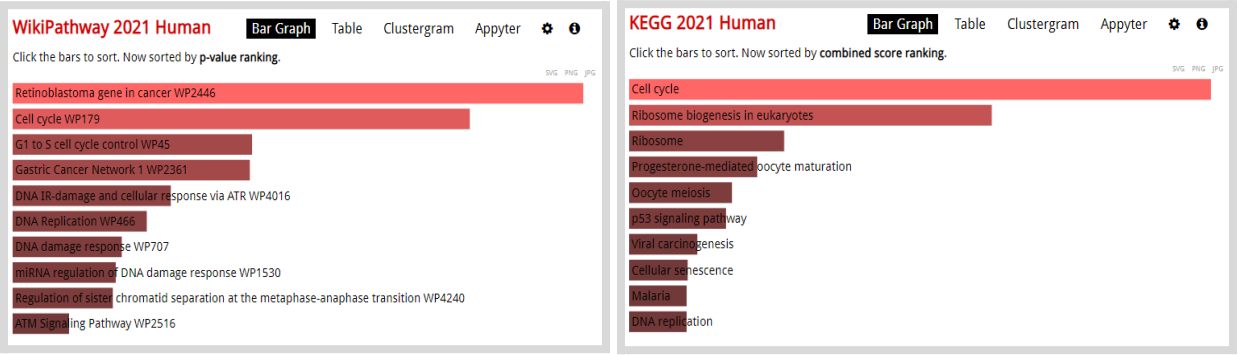

Figure 7. **Enrich R Results:** The DEGs were inputted into Enrich R to find the relevant pathways. Using WikiPathway and KEGG, the pathways with the lowest p-values were found.

Classification:

After preprocessing and filtering our data, we trained and tested our different classification models on all of the remaining genes. The table below lays out the precision, recall, and F1-score for each of the different classification models.

	KNN Classifier	Decision Tree Model	SVM
Precision	0.542	0.645	0.812
Recall	0.548	0.643	0.770
F1-score	0.530	0.634	0.770

Table 2. **Classifier Results.** The precision, recall, and F1-score of each classification model when training on all genes.

Since our SVM performed the best in all three metrics, we filtered the dataset based on the previously found differentially expressed genes and recreated our SVM. Table 3 outlines the new results once the non-differentially expressed genes were filtered out and the SVM was retrained. Table 4 compares the retrained SVM model against the Overmyer classifier and outlines.

	All Genes	Differentially Expressed Genes
Precision	0.812	0.75
Recall	0.77	0.722
F1-score	0.77	0.722

Table 3. SVM Results. The results after retraining our SVM with only the differentially expressed genes.

	DEG SVM	Overmyer Extra Tree Model
Precision	0.75	0.917
Recall	0.722	0.917
F1-score	0.722	0.917

Table 4. Comparing our classifier with the classifier from the Overmyer study. The precision, recall, and f1-scores of our trained SVM and the Extra Tree Model were created in the Overmyer study.

Lastly, we relabeled our data according to the patient's Covid status, retrained the model using only the differentially expressed genes, and compared the results to the Overmyer study's classifier (Table 5).

	DEG-only Covid SVM	Overmyer Extra Tree Model
Precision	0.923	0.917
Recall	0.905	0.917

F1-score	0.910	0.917

Table 5. Comparing our Covid classifier with the classifier from the Overmyer study. The precision, recall, and f1-scores of our relabeled SVM and the Extra Tree Model created in the Overmyer study.

Discussion:

Principal Component Analysis:

When analyzing each of the PCA figures (2,3,4), we can see that there is some general clustering in each figure but with some being more significant than others. When conducting the PCA on the top 25% varying genes, we can see that there is some slight clustering with the Covid-19 status and the ICU status but cannot attribute much to this data because of the mere 15.0% explained variation of PC1. When analyzing the PCA conducted with all four combinations of status, we can see that the four combinations each form a rough quadrant of the plane with little overlap at 15.0% explained variation.

Classification:

When analyzing the results of the three classification models based on all 16,000 filtered genes, we can see that the SVM performed significantly better than the other two models and the Decision Tree Model was slightly better than the KNN model (Figure 8). While it can be difficult to explain why the SVM performed better than the Decision Tree Model, the KNN likely performed the worst due to an untuned k as well as the large number of dimensions (unique genes).

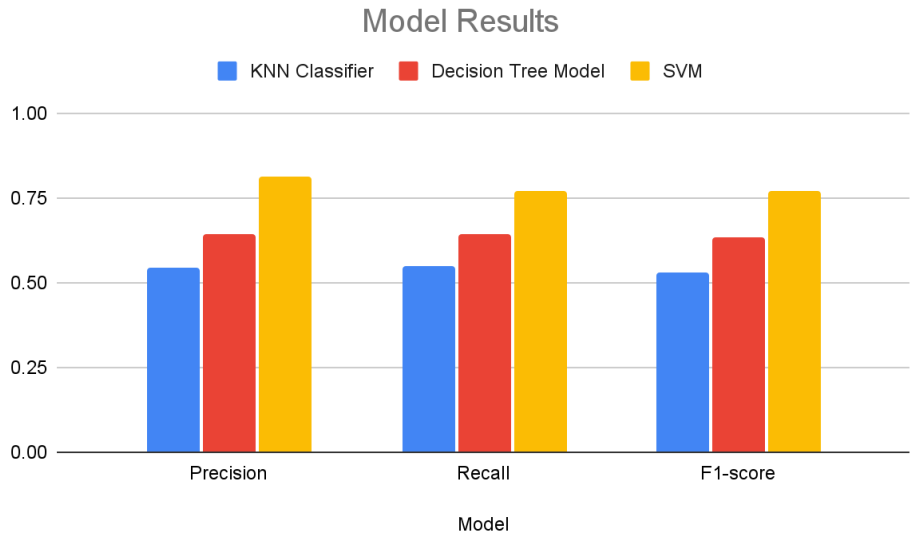

Figure 8. Comparing Results of Classifiers. A bar chart comparing the precision, recall, and F1-score of each of the three classification models. The SVM performed the best, with the decision tree model following and the KNN Classifier performed the worst.

When comparing the SVM model retrained on only the differentially expressed genes to the one trained on all the genes, we notice that the retrained model performed slightly worse (Figure 9). The retrained model might be performing worse than expected due to overfitting in the first model. While we performed 6-fold cross validation to reduce overfitting, the initial model with all 16,000 genes may have overfit the model based on differences in the non-differentially expressed genes. Thus, once these genes were filtered out, the retrained model could not take advantage of these genes and overfit them.

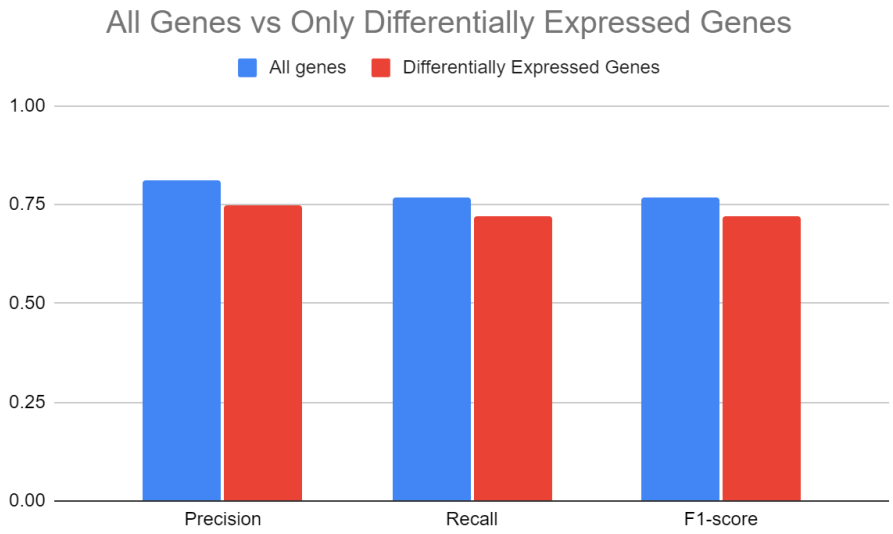

Figure 9. **Comparison of the SVM models.** A bar chart comparing the precision, recall, and F1-score of the SVM trained on all the genes and the SVM trained on only differentially expressed genes. The SVM trained on all genes performed slightly better in all three metrics.

When comparing our refined SVM classifier against the Extra Tree Model from the Overmyer study, we see a significant difference in the performance (Figure 10).

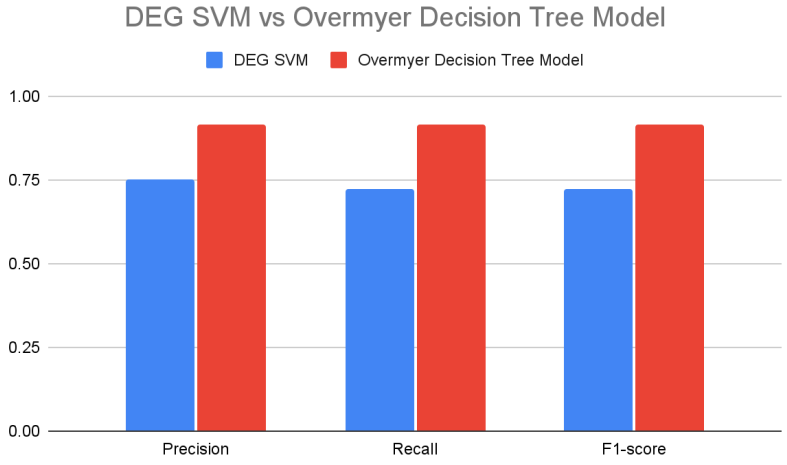

Figure 10. Comparison of our SVM model to the model from the Overmyer study. The bar graph compares the SVM trained on only differentially expressed genes to the Decision Tree model created in the Overmyer study.

The main reason this would occur is the difference in the classification type of the two models. Our SVM attempts to label each patient into one of four categories based on the Covid and ICU status of the patient. The Overmyer study on the other hand, only classifies the Covid status of a patient. Once we relabeled the data only based on Covid status and retrained the SVM model using only the differentially expressed genes, our results more closely matched the performance of the Overmyer study (Figure 11).

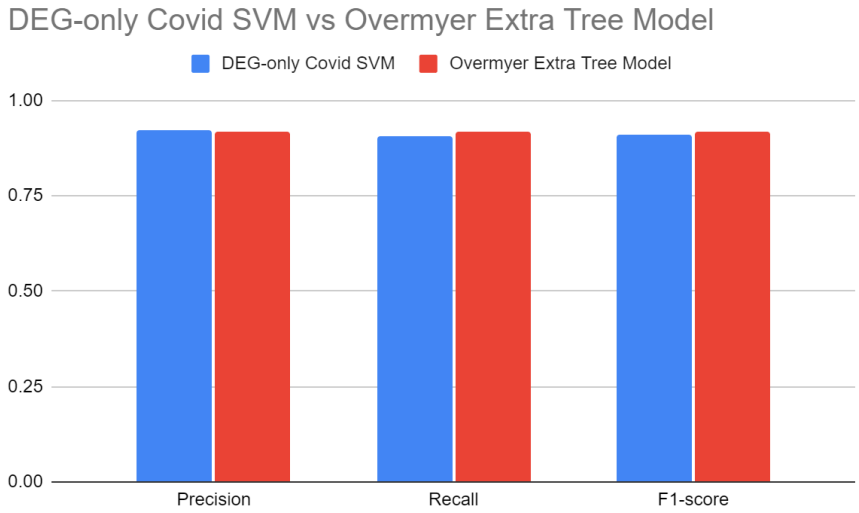

Figure 11. Covid SVM classifier compared to the Overmyer classifier. A bar graph comparing the various metrics between our Covid SVM classifier using only differentially expressed genes and the Extra Trees model created in the Overmyer classifier.

Differential Gene Expression:
Using the absolute log2FC cutoff of >1.5 and a p-value <0.05, we were able to locate 335 differentially expressed genes. From the heatmap, we can see that although there is not a complete separation, there is a general trend with the COVID samples on the left and mostly Non-COVID samples on the right. This means that although there was no perfect separation between COVID and Non-COVID samples, there was a general trend that could be observed. When we combine the differential genes located through the heatmap with a PCA (figure 6), we can observe clearer clustering within the four groups. These results are also more substantial than the previous PCAs due to the explained variation with PC1 being 22.1%.

Enrich R Results:
The Enrich R results of the DEGs had many prominent results. For example, there were many pathways regarding the cell cycle, including G1 to S cycle control, WP179 cell cycle pathway,

and cellular senescence, which involves cell death (von Kobbe, 2018), often to combat infecting viruses like COVID-19. Another portion of interesting pathways had to do with viral carcinogenesis, which viruses use to hijack the cellular machinery to facilitate their own rapid reproduction. The p53 pathway is also a prominently featured pathway in COVID research as it inhibits the replication of COVID-19 (Gorskaya, 2021). However, there were many genes that did not seem to have much relevance to COVID-19. These may eventually have significance, but there may be some confounding variables in the results, like in the case of Oocyte meiosis, so further research and analysis can be done. Pathways relating to Malaria have also been differentially expressed. For example, In malaria-endemic regions, COVID-19 has led to difficulties due to the two viruses sharing similar symptoms (Chanda-Kapata, 2020).

Limitations and Future Steps:
With reference to the limitations we faced it is pertinent to address the fact that our data only contains samples from patients who reported being affected by respiratory illnesses or demonstrated COVID-19 compatible symptoms. Hence for a future investigation, it would be optimal to incorporate healthy controls into the data, providing a more accurate representation of the results we gathered. This also ties into another limitation which focuses on the imbalance of our samples. The number of COVID samples was much larger in proportion to the number of non-COVID samples. For instance, we took into consideration a total of 100 COVID samples versus 26 for non-COVID. This disparity could potentially lead to different and inaccurate results since the data gathered would favor COVID samples, and on a greater scale, it would not represent the population as a whole or the reality of the situation. Therefore in a future analysis we would have a larger dataset that has a better ratio of COVID samples to non-COVID samples.

In addition, we would also have to consider other factors that could affect the ICU status of the patients, such as age and other preexisting conditions, like immunocompromised patients or those suffering from long-term respiratory disorders like sleep apnea or chronic obstructive pulmonary disease. This leads to the possible consideration of other elements that could influence the results, for example, one of our enriched pathways was “Progesterone Mediated Oocyte Maturation”, which is a pathway directly related to sex; therefore in a future analysis, we could account for the sex of the patient.

References

Chanda-Kapata, Pascalina et al. “COVID-19 and malaria: A symptom screening challenge for malaria endemic countries.” International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases vol. 94 (2020): 151-153. doi:10.1016/j.ijid.2020.04.007

Gorskaya Yu F, Semenova EN, Nagurskaya EV, Bekhalo VA, Nesterenko VG (2021) Apoptosis and P53 Activation are involved in COVID-19 Pathogenesis. COVID-19 Pandemic: Case Studies & Opinions 02(02): 208–213.

Guan WJ, Ni ZY, Hu Y, … , Zhong NS; China Medical Treatment Expert Group for Covid-19. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020 Apr 30;382(18):1708-1720. doi: 10.1056/NEJMoa2002032. Epub 2020 Feb 28. PMID: 32109013; PMCID: PMC7092819.

Overmyer KA, Shishkova E, Miller IJ, …, Jaitovich A. Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst. 2021 Jan 20;12(1):23-40.e7. doi: 10.1016/j.cels.2020.10.003. Epub 2020 Oct 8. PMID: 33096026; PMCID: PMC7543711.

Pedregosa, F. κ.ά. ‘Scikit-learn: Machine Learning in Python’. Journal of Machine Learning Research 12 (2011): 2825–2830. Print.

Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, ..., Karlsen TH. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020 Oct 15;383(16):1522-1534. doi: 10.1056/NEJMoa2020283. Epub 2020 Jun 17. PMID: 32558485; PMCID: PMC7315890.

von Kobbe, C. (2018). Cellular senescence: a view throughout organismal life. Cell. Mol. Life Sci. 75, 3553–3567. doi: 10.1007/s00018-018-2879-8

WHO COVID-19 Dashboard. Geneva: World Health Organization, 2020. Available online: https://covid19.who.int/ (last cited: Apr 28 2022).