International Journal of
Molecular Sciences

MDPI

Article

Classification of COVID-19 Patients into Clinically Relevant Subsets by a Novel Machine Learning Pipeline Using Transcriptomic Features

Andrea R. Daamen [1,2], Prathyusha Bachali [1,2], Amrie C. Grammer [1,2] and Peter E. Lipsky [1,2,*]

[1] AMPEL BioSolutions LLC, Charlottesville, VA 22902, USA
[2] RILITE Research Institute, Charlottesville, VA 22902, USA
* Correspondence: peterlipsky@comcast.net

Abstract: The persistent impact of the COVID-19 pandemic and heterogeneity in disease manifestations point to a need for innovative approaches to identify drivers of immune pathology and predict whether infected patients will present with mild/moderate or severe disease. We have developed a novel iterative machine learning pipeline that utilizes gene enrichment profiles from blood transcriptome data to stratify COVID-19 patients based on disease severity and differentiate severe COVID cases from other patients with acute hypoxic respiratory failure. The pattern of gene module enrichment in COVID-19 patients overall reflected broad cellular expansion and metabolic dysfunction, whereas increased neutrophils, activated B cells, T-cell lymphopenia, and proinflammatory cytokine production were specific to severe COVID patients. Using this pipeline, we also identified small blood gene signatures indicative of COVID-19 diagnosis and severity that could be used as biomarker panels in the clinical setting.

Keywords: COVID-19; severity; classification; transcriptomics; bioinformatics; machine learning

check for updates

Citation: Daamen, A.R.; Bachali, P.; Grammer, A.C.; Lipsky, P.E. Classification of COVID-19 Patients into Clinically Relevant Subsets by a Novel Machine Learning Pipeline Using Transcriptomic Features. *Int. J. Mol. Sci.* **2023**, *24*, 4905. https://doi.org/10.3390/ijms24054905

Academic Editor: Andrzej Kloczkowski

Received: 15 January 2023
Revised: 24 February 2023
Accepted: 1 March 2023
Published: 3 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Since its emergence in December 2019, infection by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), has led to the deaths of over 6.5 million individuals worldwide [1–3]. COVID-19 is characterized by a wide range of disease presentations from mild with flu-like symptoms to severe with acute hypoxic respiratory failure (AHRF) requiring hospitalization and admittance to the intensive care unit (ICU) [4]. In addition, symptoms may persist for months after patients present with acute illness [5]. A number of pre-existing clinical risk factors (age, gender, obesity, respiratory conditions, diabetes, immunodeficiency) [6,7] and post-infection multi-organ disease complications, including cardiac injury, thrombosis, renal disease, and liver injury [8–10], have been associated with severe COVID. Despite intense study, however, the major drivers of severe and potentially fatal COVID-19 have not been fully elucidated.

Considerable resources have been applied to study the immune response to SARS-CoV-2 infection as a means to understand COVID-19 pathogenesis in greater detail [11,12]. Among these efforts, our group and others have employed a bioinformatics-based approach using multi-omic, and in particular, transcriptomic data to construct immune profiles of COVID-19 patients at various stages of disease onset and severity [13–22]. These studies have identified numerous abnormalities in the strength and nature of the adaptive and innate immune responses of severe COVID patients, including neutrophilia, T-cell lymphopenia, plasmablast expansion, antibody or autoantibody production, and IFN levels. With this accumulated knowledge base, the focus has evolved to attempt to apply the information to identify subsets of individuals with COVID-19 with prognostic relevance.

Recent advances in computer science and the use of artificial intelligence (AI) and machine learning (ML) for clinical applications offer a promising approach to identify biomarkers and predict the risk of SARS-CoV-2 infected individuals to develop severe and possibly fatal disease [23,24]. Several recent studies have employed ML or deep learning approaches to patient demographic data [25], lung CT scan images [26–28], and transcriptomic data [29–33] to stratify COVID-19 patients with varying degrees of success. However, more work is required to improve model performance and reproducibility before biomarkers of severe COVID-19 identified by ML can be translated into a clinical setting. This is dependent on: (1) developing a standardized approach to dealing with high dimensional data; (2) selecting the best ML algorithm for each application; and (3) selecting the most informative features as input for ML applications. To that end, we have developed an iterative ML pipeline based on curated gene signatures previously used to characterize immune profiles of COVID-19 patients from whole blood gene expression data [13,14]. As a result, we have identified immunological signatures and individual genes with strong predictive capacity for classifying COVID-19 patients most at risk of severe disease that could be employed as a diagnostic and prognostic tool in the clinic.

2. Results

2.1. A Gene Expression-Based Iterative Machine Learning Approach to Classification of COVID-19 Patients

We employed three publicly available whole blood transcriptome datasets (GSE161731, GSE172114, and PRJNA777938) that had previously been analyzed for immune profiling of COVID-19 patients based on gene signature enrichment by gene set variation analysis (GSVA) [14]. The combined datasets consisted of individuals with COVID-19, individuals with non-COVID-induced AHRF, individuals admitted to the ICU without AHRF, and healthy controls. COVID-19 patients were further categorized as "non-critical" if they exhibited symptoms but were not hospitalized or were admitted to a non-critical care ward and as "critical" if they exhibited more severe symptoms requiring hospitalization and admittance to the ICU.

To determine the top immune signatures and genes differentiating subsets of COVID-19 patients from healthy individuals and other ICU patients, combined GSVA enrichment scores for 40 curated immune cell and pathway gene signatures [13,14] were used as features to train 9 ML algorithms (Figure 1, Table S1). Each ML algorithm was used for 4 binary classifications: COVID patients versus healthy individuals, non-critical COVID patients versus healthy individuals, critical versus non-critical COVID patients, and COVID ICU patients versus non-COVID ICU patients. Then, the top 5 performing ML algorithms were employed in an iterative approach to identify the GSVA modules contributing most to each classification (Figure 1). After each iteration, feature importance was calculated for each ML algorithm and the top 50% of features were used for the next round of ML. Then, the $\log_2$ expression values of genes composing the final 10 gene modules were used as features for ML algorithms to identify the top 20 individual genes that could effectively classify COVID-19 patients in each comparison. Expression values used as input for the final ML iteration were normalized to account for batch effects between separate RNA-seq datasets (Figures S1–S4A). To ensure reproducibility of the results, each ML iteration was repeated 10 times and the most represented modules or gene features were used moving forward. Overall, the ML algorithms successfully performed each binary classification and the top performing algorithms for each comparison achieved accuracies of >0.7 for all iterations of ML. In addition, accuracies improved between the first and last ML iteration as the gene modules were refined and distilled into individual gene features, and final accuracies reached >0.9 for the top 5 algorithms in each classification (Table S2). Notably, the final gene lists derived from this iterative ML approach reflected distinct immune cell and inflammatory pathways providing the greatest contribution to each binary classification of COVID-19 patients (Table S3).

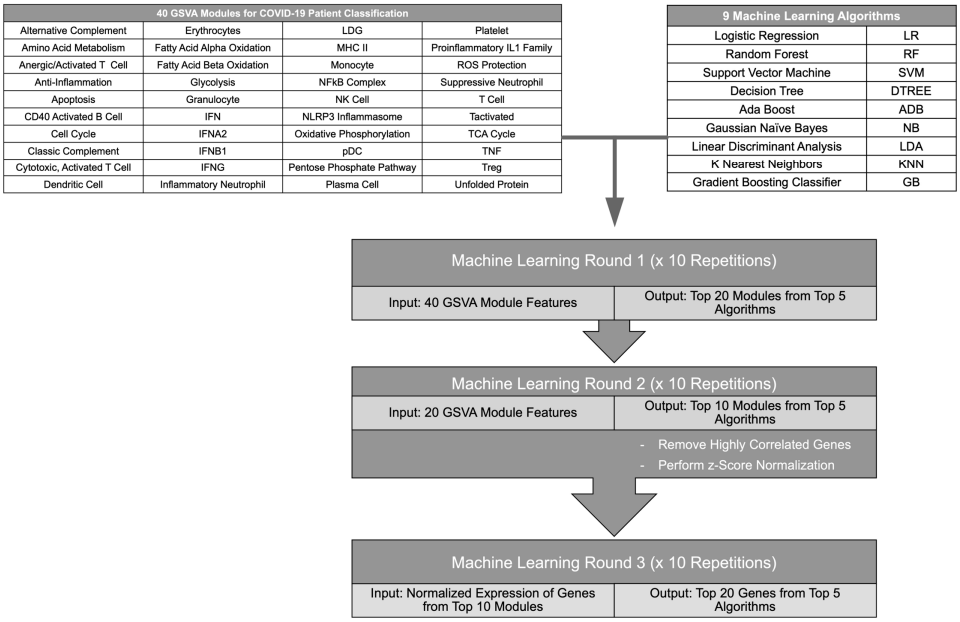


Figure 1. Flow diagram of the iterative ML pipeline. Graphic depicting steps in the iterative ML pipeline used for classification of COVID-19 patients. GSVA scores of gene modules representing immune cells and pathways were used as input for 9 different ML algorithms in 2 iterations after which a final ML iteration used z-score-normalized gene expression values. The flow chart was created in Microsoft Powerpoint.

2.2. Machine Learning Models Reveal Genes Critical for Classification of COVID-19 Patients from Healthy Individuals

The iterative ML pipeline was first applied to determine the top modules and individual genes for classification of COVID-19 patients from healthy individuals. For the initial iteration, support vector machine (SVM) was consistently the top performing algorithm with average areas under the receiver operating characteristic and precision/recall curves (AU-ROC and AU-PR) of 1.0 (Figure 2A) as well as an average overall accuracy of 1.0, indicative of perfect model performance across 10 repetitions (Table S2). The 10 modules with the highest feature importance were divided between pathways related to cell proliferation/metabolism (Cell Cycle, Pentose Phosphate Signature, Glycolysis, Oxidative Phosphorylation (OxPhos), Fatty Acid Alpha Oxidation (FAAO)) and inflammatory modules (Anti inflammation, Alternative Complement Pathway, Granulocyte, MHC II, Proinflammatory IL-1 Family) (Figure 2B, Table S3). In the final ML iteration, the perfect classification metrics observed in the first iteration were maintained, but linear regression (LR) and random forest (RF) were the top performing algorithms (Figure 2C, Table S2). This result highlights the importance of comparing multiple algorithms when inputting different sets of features to the ML pipeline. The final 20 genes with greatest performance in classifying COVID-19 patients from healthy individuals were dominated by genes from modules of the Cell Cycle (CCNB1, CCNB2, CDC20, CEP55, GINS2, MCM10, MKI67, PTTG1) and OxPhos (COX6C, COX7B, COX7C, NDUFAF5, NDUFAF7, NDUFB3, TMEM126B, UQCRB) but also included genes in the Glycolysis (SLC2A3), and Granulocyte (CD177, CLC, OSM) modules (Figures 2D and S1B, Table S3). To determine the exact relationship between expression of these 20 genes and COVID-19 classification, we used SHapley Additive exPlanations (SHAP) values, which provide information on the magnitude and directionality of the contribution of each gene to classification of an individual as a COVID-19 patient or a healthy control. SHAP values are displayed in a summary plot (Figure 2E) in which each dot represents the expression of each gene in an individual patient and higher gene expression values in conjunction with increased SHAP values are indicative of a positive association with COVID-19 classification, and lower gene expression values in conjunction with increased SHAP values indicate a negative correlation with COVID-19

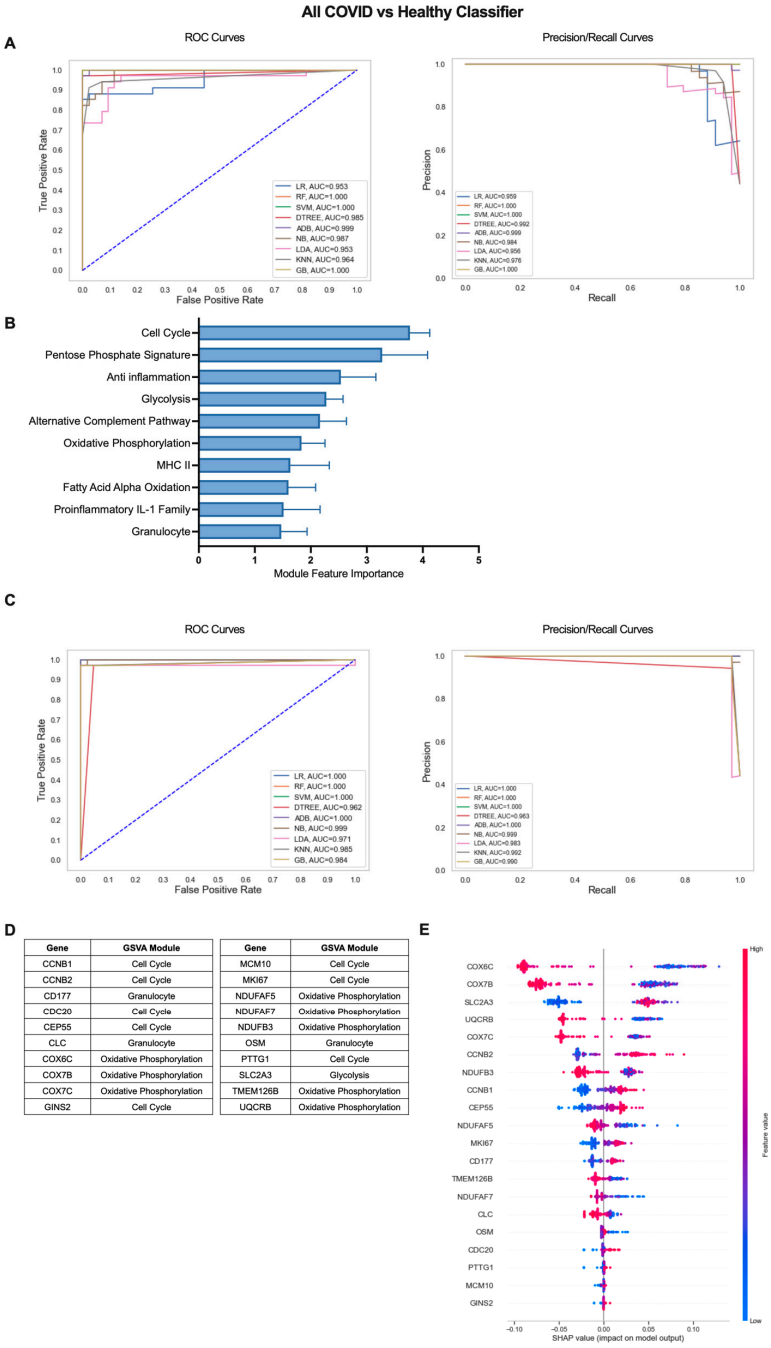
classification. Overall, we found that increased expression of cell cycle and glycolysis genes and decreased expression of OxPhos genes were associated with COVID-19 classification (Figure 2E). Genes in the Granulocyte module were split with increased expression of CD177 and decreased expression of CLC and OSM contributing to COVID classification (Figure 2E).

The COVID-19 patient cohort consisted of individuals with mild, non-critical disease as well as individuals with severe disease requiring ICU admission. Therefore, we sought to determine whether an iterative ML approach could also effectively identify subtle differences between COVID-19 patients with mild disease and healthy individuals (Figure 3). Interestingly, the non-critical COVID vs. healthy classifier performed similarly to classification including COVID-19 patients with severe disease. In the first ML iteration, SVM was the top performing algorithm over 10 repetitions with an average AU-ROC of 0.994, AU-PR of 0.992, and accuracy of 0.976 (Figure 3A, Table S2). The top 10 gene modules for the non-critical COVID-19 vs. healthy classifier shared common proliferation/metabolism pathways with results when including all COVID-19 patients but, interestingly, had more representation of immune cell gene modules including Plasma Cells (PCs), CD40 Activated B Cells, and Inflammatory Neutrophils (Figure 3B, Table S3). Algorithm performance improved by the final ML iteration and LR and RF algorithms performed the best with perfect classification metrics in 9 out of 10 repetitions (Figure 3C, Table S2). The additional immune modules selected as top features identifying non-critical COVID-19 patients from healthy controls were reflected in the final list of genes as in addition to Cell Cycle (CCNB1, CCNB2, CCNE1, CDC20, CEP55, E2F3) and OxPhos (COX7B, COX7C, DNAJC15, NDUFA5, NDUFAF7, TMEM126B, UQCRB), genes in the Inflammatory Neutrophil (CARD16, CFL1, CLEC4D, HBC21), CD40 Activated B Cell (DUSP4), Alternative Complement Pathway (CFD), and Apoptosis (FAS) modules were also included (Figures 3D and S2B, Table S3). SHAP values for the top 20 genes indicated that increased expression of genes in the Cell Cycle, Alternative Complement Pathway, CD40 Activated B Cell, and Inflammatory Neutrophil modules and decreased expression of genes in the OxPhos and apoptosis modules were associated with classification of non-critical COVID-19 patients as compared to healthy individuals (Figure 3E). Thus, an iterative ML approach effectively differentiated COVID-19 patients with varying levels of disease severity from healthy individuals based solely on expression of gene modules and individual genes in each patient.

2.3. Iterative Machine Learning Effectively Predicts Disease Severity of COVID-19 Patients Based on Gene Expression

To extend the clinical applications of gene expression-based classification of COVID-19 patients, we employed the iterative ML pipeline to the classification of COVID-19 patients with critical, severe disease requiring hospitalization from patients with a mild, non-critical form of the disease (Figure 4). For the critical vs. non-critical COVID-19 classifier, the gradient boosting (GB) algorithm performed best in the first iteration with an average AUC-ROC and AUC-PR of 0.82 and an average accuracy of 0.83 across 10 repetitions (Figure 4A, Table S2). The top 10 gene modules distinguishing critical from non-critical COVID-19 largely consisted of inflammatory cell/pathway modules and, in particular, T cells (T Cell, Inflammatory Neutrophil, IFN, Tactivated, Treg, Cytotoxic, Activated T Cell) as well as metabolic pathways (Amino Acid (AA) Metabolism, Pentose Phosphate Signature, FAAO), and the stress response (Unfolded Protein) (Figure 4B, Table S3). As with the COVID-19 vs. healthy classifier, performance across all algorithms improved by the final iteration and GB remained the top performer with an average AUC-ROC of 0.98, AUC-PR of 0.98, and overall accuracy of 0.93 (Figure 4C, Table S2). As a result, the majority of the top 20 individual genes identifying critical COVID-19 patients with more severe disease were derived from the Inflammatory Neutrophil module (CARD16, CAST, CD177, FKBP5, H2BC21, HIF1A, MCEMP1, NT5C3A) as well as genes from the IFN (EIF2AK2, SP100), Unfolded Protein (DERL1, ERAP1, SSR1), AA Metabolism (OXCT1), and T-cell-related modules (CCR2, CD28, CREM, EOMES, SGK1) (Figures 4D and S3B,

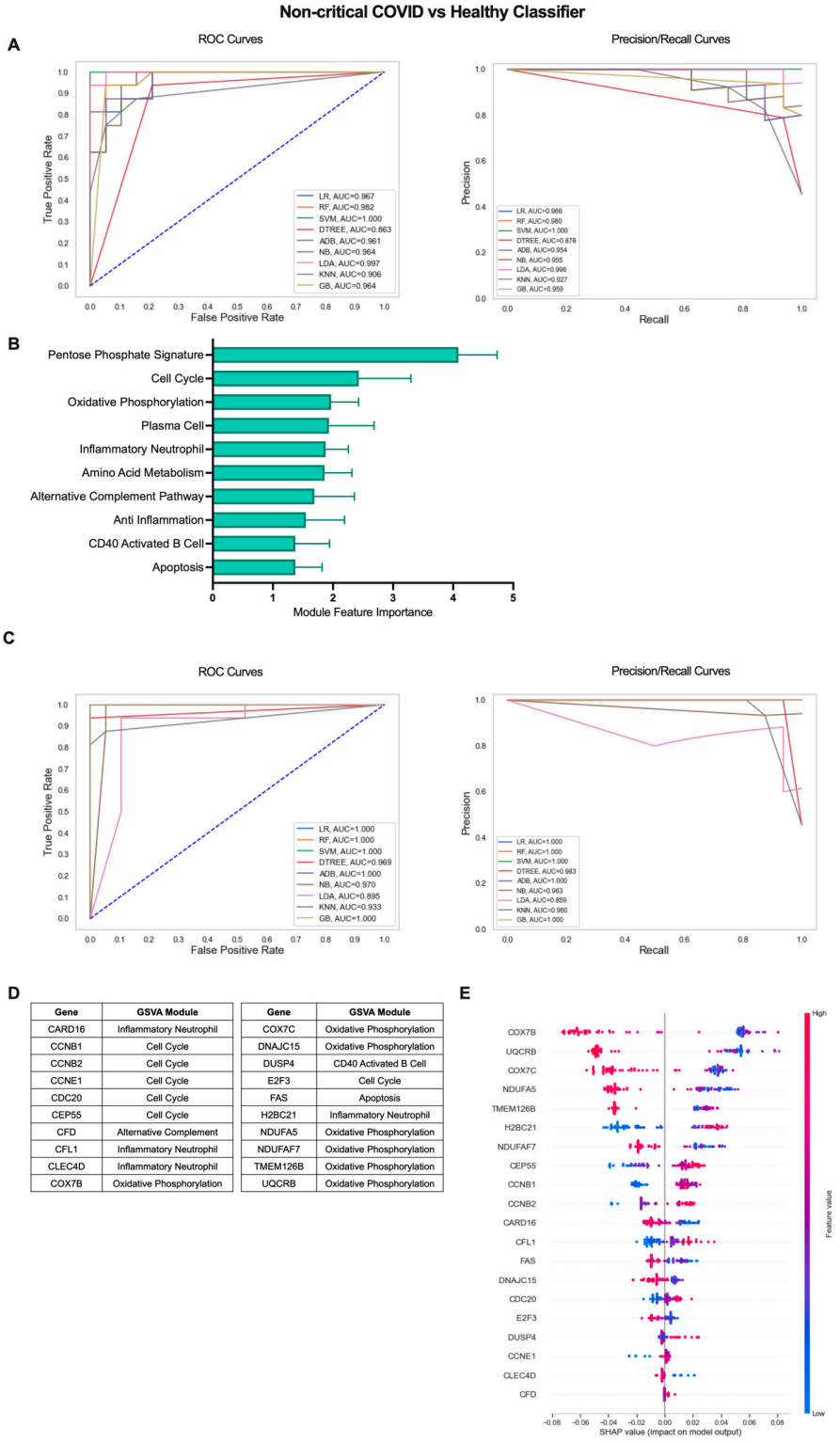
Table S3). SHAP analysis of the directionality of relationships between relative expression of the top 20 genes and classification of critical COVID-19 revealed that, as a whole, critical classification was associated with increased expression of inflammatory neutrophil and IFN genes and decreased expression of T-cell, metabolism, and stress response genes (Figure 4E). These results emphasize the differences in the nature of inflammation present in critical COVID-19 patients as compared to those with mild disease.



Gene	GSVA Module
CCNB1	Cell Cycle
CCNB2	Cell Cycle
CD177	Granulocyte
CDC20	Cell Cycle
CEP55	Cell Cycle
CLC	Granulocyte
COX6C	Oxidative Phosphorylation
COX7B	Oxidative Phosphorylation
COX7C	Oxidative Phosphorylation
GINS2	Cell Cycle

Gene	GSVA Module
MCM10	Cell Cycle
MKI67	Cell Cycle
NDUFAF5	Oxidative Phosphorylation
NDUFAF7	Oxidative Phosphorylation
NDUFB3	Oxidative Phosphorylation
OSM	Granulocyte
PTTG1	Cell Cycle
SLC2A3	Glycolysis
TMEM126B	Oxidative Phosphorylation
UQCRB	Oxidative Phosphorylation

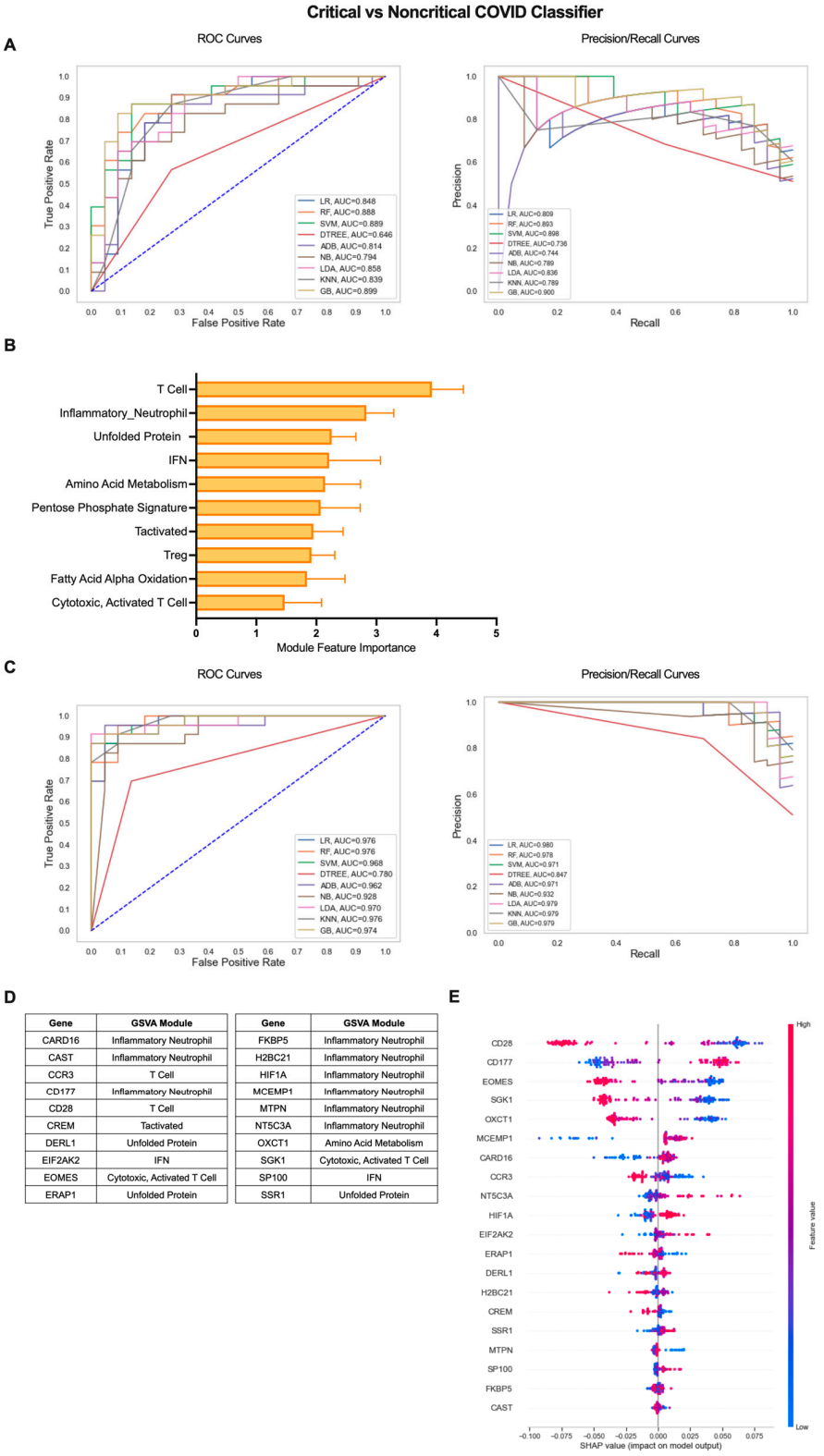
Figure 2. Iterative ML COVID-19 patient vs. healthy individual classifier results. (**A**) Representative ROC and PR curves from the first ML pipeline iteration for the all COVID-19 patients vs. healthy control classifier. (**B**) Bar graph of feature importance values for the top 10 gene modules resulting from the second ML pipeline iteration. (**C**) Representative ROC and PR curves from the third ML pipeline iteration for the all COVID-19 patients vs. healthy control classifier. (**D**) Top 20 gene features resulting from the third ML pipeline iteration. (**E**) SHAP summary plot depicting the magnitude and directionality of relationships between the top 20 gene features and COVID-19 classification.



Gene	GSVA Module
CARD16	Inflammatory Neutrophil
CCNB1	Cell Cycle
CCNB2	Cell Cycle
CCNE1	Cell Cycle
CDC20	Cell Cycle
CEP55	Cell Cycle
CFD	Alternative Complement
CFL1	Inflammatory Neutrophil
CLEC4D	Inflammatory Neutrophil
COX7B	Oxidative Phosphorylation

Gene	GSVA Module
COX7C	Oxidative Phosphorylation
DNAJC15	Oxidative Phosphorylation
DUSP4	CD40 Activated B Cell
E2F3	Cell Cycle
FAS	Apoptosis
H2BC21	Inflammatory Neutrophil
NDUFA5	Oxidative Phosphorylation
NDUFAF7	Oxidative Phosphorylation
TMEM126B	Oxidative Phosphorylation
UQCRB	Oxidative Phosphorylation

Figure 3. Iterative ML non-critical COVID-19 patient vs. healthy individual classifier results. (**A**) Representative ROC and PR curves from the first ML pipeline iteration for the non-critical COVID-19 patient vs. healthy control classifier. (**B**) Bar graph of feature importance values for the top 10 gene modules resulting from the second ML pipeline iteration. (**C**) Representative ROC and PR curves from the third ML pipeline iteration for the non-critical COVID-19 patient vs. healthy control classifier. (**D**) Top 20 gene features resulting from the third ML pipeline iteration. (**E**) SHAP summary plot depicting the magnitude and directionality of relationships between the top 20 gene features and non-critical COVID-19 classification.



Gene	GSVA Module
CARD16	Inflammatory Neutrophil
CAST	Inflammatory Neutrophil
CCR3	T Cell
CD177	Inflammatory Neutrophil
CD28	T Cell
CREM	Tactivated
DERL1	Unfolded Protein
EIF2AK2	IFN
EOMES	Cytotoxic, Activated T Cell
ERAP1	Unfolded Protein

Gene	GSVA Module
FKBP5	Inflammatory Neutrophil
H2BC21	Inflammatory Neutrophil
HIF1A	Inflammatory Neutrophil
MCEMP1	Inflammatory Neutrophil
MTPN	Inflammatory Neutrophil
NT5C3A	Inflammatory Neutrophil
OXCT1	Amino Acid Metabolism
SGK1	Cytotoxic, Activated T Cell
SP100	IFN
SSR1	Unfolded Protein

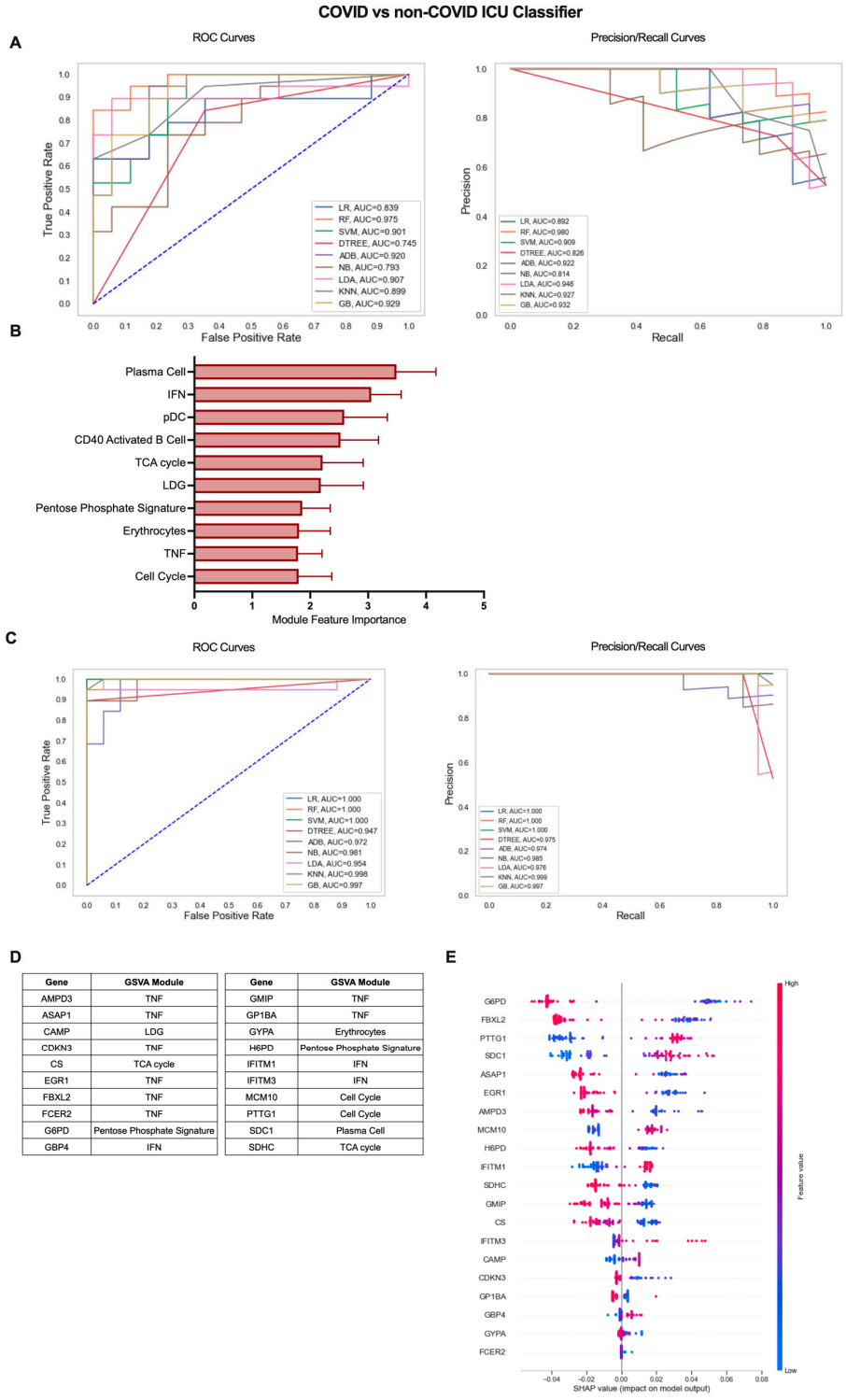
Figure 4. Iterative ML critical vs. non-critical COVID-19 patient classifier results. (**A**) Representative ROC and PR curves from the first ML pipeline iteration for the critical vs. non-critical COVID-19 patient classifier. (**B**) Bar graph of feature importance values for the top 10 gene modules resulting from the second ML pipeline iteration. (**C**) Representative ROC and PR curves from the third ML pipeline iteration for critical vs. non-critical COVID-19 patient classifier. (**D**) Top 20 gene features resulting from the third ML pipeline iteration. (**E**) SHAP summary plot depicting the magnitude and directionality of relationships between the top 20 gene features and critical COVID-19 classification.

2.4. A Gene Expression-Based Machine Learning Approach Identifies Genes Distinguishing COVID-19 ICU Patients from Other Patients Admitted to the ICU

Our previous analysis of patients admitted to the ICU with or without COVID-19-induced AHRF found few differences in gene-expression-based immune profiles, predominantly centered on the nature and magnitude of the PC antibody response [14]. To explore these differences in greater depth, we utilized the iterative ML pipeline for classification of COVID-19 vs. non-COVID-19 ICU patients (Figure 5). The first iteration of this classifier had the weakest initial algorithm performance, and SVM performed best in 5 out of 10 repetitions with an average AUC-ROC of 0.77, AUC-PR of 0.81, and accuracy of 0.76 (Figure 5A, Table S2). Of the top 10 gene modules for the COVID-19 vs. non-COVID-19 ICU classifier, the PC module had the greatest overall feature importance, as expected from our previous work (Figure 5B, Table S3). The remaining modules in the top 10 included those related to the inflammatory response (IFN, pDC, CD40 Activated B Cell, LDG, Erythrocytes, TNF) as well as cell proliferation and metabolism (Cell Cycle, TCA Cycle, Pentose Phosphate Signature). Algorithm performance improved dramatically by the final ML iteration in the pipeline, and linear regression (LR) was the best overall algorithm with average AUC-ROC of 0.91, AUC-PR of 0.99, and accuracy of 0.97 (Figure 5C, Table S2). The top 20 genes predominantly originated from the TNF module (AMPD3, ASAP1, CDKN3, EGR1, FBXL2, FCER2, GMIP, GP1BA) but also included genes from the PC (SDC1), IFN (GBP4, IFITM1, IFITM3), Cell Cycle (MCM10, PTTG1), TCA cycle (CS, SDHC), Pentose Phosphate Signature (G6PD, H6PD), LDG (CAMP), and Erythrocyte (GYPA) modules (Figures 5D and S4B, Table S3). Interestingly, the SHAP summary plot indicated that decreased expression of most of these genes including those from the TNF and metabolism-related modules and increased expression of genes from the PC, IFN, and LDG modules was associated with COVID-19 classification from all patients admitted to the ICU (Figure 5E). Altogether, these results demonstrate that gene expression data can effectively classify subsets of COVID-19 patients and provide insight into inflammatory cells and pathways with the greatest contribution to disease severity.

2.5. Validation of the Iterative ML Pipeline in an Independent COVID-19 Patient Dataset

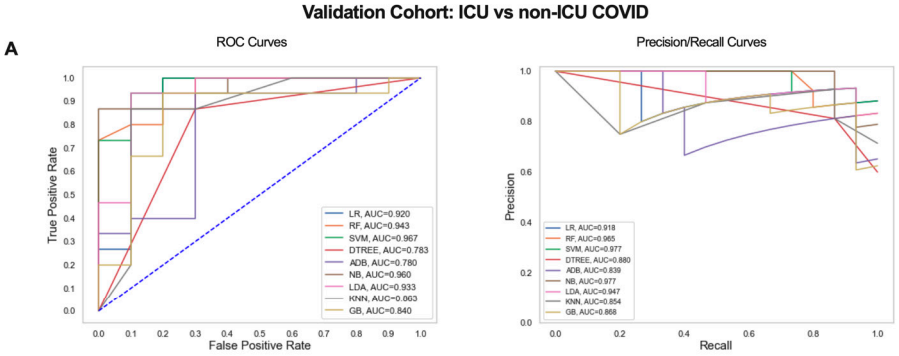
Next, we determined whether the small gene signature output from our iterative ML pipeline could distinguish COVID-19 patients with mild/moderate and severe disease in an independent dataset (GSE157103) [21]. Blood RNA-seq data from 100 hospitalized COVID-19 patients (50 admitted to the ICU and 50 non-ICU) was used to validate the 20 gene signature from the critical vs. non-critical COVID-19 classifier (Figure 4D, Table S3). Scaled $\log_2$ expression values for the 20 genes were used as input for 9 ML algorithms and combined ROC and PR curves were generated (Figure 6A). SVM was the top performing algorithm with excellent AUC-ROC and AUC-PRs of 0.97. Classification performance metrics were high for all algorithms with sensitivities ranging from 0.67 to 0.87, specificities from 0.7 to 1, and accuracies from 0.72 to 0.88 (Figure 6B). Therefore, gene signatures derived from the iterative ML pipeline can be successfully applied to classification of COVID-19 severity in an independent patient cohort and achieve a comparable level of model performance.



Gene	GSVA Module
AMPD3	TNF
ASAP1	TNF
CAMP	LDG
CDKN3	TNF
CS	TCA cycle
EGR1	TNF
FBXL2	TNF
FCER2	TNF
G6PD	Pentose Phosphate Signature
GBP4	IFN

Gene	GSVA Module
GMIP	TNF
GP1BA	TNF
GYPA	Erythrocytes
H6PD	Pentose Phosphate Signature
IFITM1	IFN
IFITM3	IFN
MCM10	Cell Cycle
PTTG1	Cell Cycle
SDC1	Plasma Cell
SDHC	TCA cycle

Figure 5. Iterative ML COVID-19 vs. non-COVID-19 ICU patient classifier results. (**A**) Representative ROC and PR curves from the first ML pipeline iteration for the COVID-19 vs. non-COVID-19 ICU patient classifier. (**B**) Bar graph of feature importance values for the top 10 gene modules resulting from the second ML pipeline iteration. (**C**) Representative ROC and PR curves from the third ML pipeline iteration for the COVID-19 vs. non-COVID-19 ICU patient classifier. (**D**) Top 20 gene features resulting from the third ML pipeline iteration. (**E**) SHAP summary plot depicting the magnitude and directionality of relationships between the top 20 gene features and COVID-19 ICU classification.



B

Classifier	AUC-ROC	AUC-PR	Sensitivity	Specificity	Accuracy
SVM	0.97	0.98	0.87	0.9	0.88
LR	0.92	0.92	0.87	0.9	0.88
RF	0.94	0.96	0.73	1	0.84
NB	0.96	0.98	0.87	0.8	0.84
KNN	0.86	0.85	0.87	0.9	0.88
LDA	0.93	0.95	0.73	0.9	0.8
DTREE	0.78	0.88	0.87	0.7	0.8
GB	0.84	0.87	0.67	0.8	0.72
ADB	0.78	0.84	0.8	0.7	0.76

Figure 6. Critical vs. non-critical COVID-19 classifier gene signature validation results. (**A**) ROC and PR curves and (**B**) table of performance metrics for ML results using the 20 gene signature from Figure 4D to compare ICU and non-ICU COVID-19 patients in an independent validation cohort.

3. Discussion

The heterogeneity of COVID-19 and the wide range of individuals affected by acute and/or persistent disease manifestations necessitates an accurate and reproducible approach to identify drivers of pathology and disease trajectory over time. We have developed a novel, iterative ML approach to identify key molecular signatures stratifying COVID-19 patients based on disease severity and to differentiate them from healthy individuals and those with other infectious or non-infectious causes of AHRF. This information could be leveraged to predict whether an individual that tests positive for COVID-19 will develop mild/moderate disease warranting standard of care treatment or severe disease requiring more careful monitoring and deployment of targeted therapeutics to prevent multi-organ damage and death. Thus, our work is directly applicable to optimizing the deployment of medical resources for COVID-19 patient care. As a result, we identified four separate gene biomarker panels distinguishing COVID-19 patients from healthy individuals, critical from non-critical COVID-19 patients, and critical COVID-19 patients from those admitted to the ICU for non-COVID-19 AHRF. Furthermore, each classifier achieved excellent ML performance metrics across 9 different ML algorithms and over 10 repetitions of each binary classification and effectively translated to an independent validation cohort of COVID-19 patients.

In developing a novel ML pipeline that would yield the most reliable and reproducible results, we considered a number of critical factors including (1) the type of input data to use, (2) the ML algorithm(s) that would be most appropriate, and (3) the method to select the most informative, non-redundant features. First, we used gene expression data as input for the ML pipeline as it produces thousands of data points and, thus, provides a comprehensive overview of COVID-19 disease profiles that is easily obtainable from a single blood draw [34–36]. In addition, our previous studies demonstrated the utility of bulk transcriptome data to characterize COVID-19 immunopathology across different tissues and levels of disease activity [13,14], results that have been validated by other multi-omics studies [15–22]. Second, to account for differences and potential biases in ML algorithms used for patient stratification, we utilized 9 different supervised classifiers spanning a wide range of approaches to model construction. These included if/then (DTREE), regression (LR), ensemble (RF, ADB, GBM), Bayesian (NB), dimensionality reduction (LDA), and instance-based algorithms (SVM, KNN) [37–39]. In addition, multiple repetitions of each

step in the ML pipeline further increased confidence in the reliability and reproducibility of the results. Finally, in working with high-throughput gene expression data, we faced the issue of feature selection and how to reduce transcriptome-wide datasets into a manageable number of gene features that would optimize ML algorithm performance [40,41]. We accounted for this by using curated gene modules representative of immune cells and pathways previously associated with COVID-19 pathogenesis. Then, using an iterative approach in which the most informative features were selected moving forward allowed us to reduce the number of genes used for the final iteration and select the top 20 overall genes able to distinguish COVID-19 patients in each classifier.

This iterative ML pipeline was used for classification of COVID-19 patients from healthy individuals, critical from non-critical COVID-19 patients, and COVID-19 from non-COVID-19 patients admitted to the ICU with AHRF. Overall, the pathologic cell types and processes represented in the resulting modules and genes reflected previous characterizations of COVID-19 patients, providing validation for these results. Comparing differences in the output of each classifier provided additional insights into the distinguishing features of severe manifestations of COVID-19 as compared to other conditions. Classification of all COVID-19 patients from healthy individuals was dependent on increased expression of genes involved in the cell cycle and decreased expression of genes involved in mitochondrial function through OxPhos. Dysregulated expression of cell cycle genes in COVID-19 patients has been reported in several studies in conjunction with outgrowth of peripheral blood leukocyte populations and thus may serve as a marker of generalized inflammation in infected individuals [42–44]. The decreased expression of OxPhos genes is indicative of mitochondrial dysfunction and oxidative stress, both of which are a general feature of the response to viral infection and have also been implicated specifically in COVID-19 patients [45–47]. Interestingly, isolating non-critical COVID-19 patients in comparison to healthy individuals yielded a final gene list that, in addition to cell cycle and OxPhos genes, incorporated genes linked with pro-inflammatory cell populations, including inflammatory neutrophils and activated B cells. Our group and others have previously associated neutrophil-driven inflammation and rapid outgrowth of antibody secreting cells (ASCs) with early SARS-CoV-2 infection [14,48], suggesting that these signatures are early signs of mild/moderate disease and that continued expansion of these populations could result in critical illness.

The final 20 gene list from the critical vs. non-critical COVID-19 classifier indicated that increased expression of inflammatory neutrophil, IFN, and stress response pathway genes in conjunction with decreased T-cell genes were hallmarks of severe COVID-19. Most of the 20 gene signature was composed of inflammatory neutrophil genes, which were originally identified in the blood of severe COVID-19 patients [49,50]. Of these, *CARD16* and *H2BC21* were shared in the signature differentiating non-critical COVID-19 patients from healthy individuals. However, their increased expression in critical patients emphasizes the greater prevalence of pathologic neutrophils in promoting severe disease complications. An elevated IFN response has also been associated with severe COVID-19 in numerous studies in line with a dysregulated systemic inflammatory response and increased expression of stress response genes as tissue damage occurs [16,51,52]. As pro-inflammatory response signatures were increased in severe COVID-19 patients, this was accompanied by a decrease in immunoregulatory population genes and, in particular, T cells indicative of T-cell lymphopenia, which has been frequently described in severe COVID-19 [15,53,54]. Notably, the specific inclusion of cytotoxic-T-cell-specific genes (*EOMES* and *SGK1*) in the 20 gene signature differentiating critical patients further supports a role for inability to clear viral infection in COVID-19 severity.

We also used our iterative ML pipeline to differentiate ICU patients with COVID-19 from those admitted with non-COVID-19 conditions. The most prevalent module for genes resulting from this classifier were indicative of a high TNF response in COVID-19 ICU patients and high expression of TNF family members has been found in damaged tissue of COVID-19 patients indicative of severe disease [55]. This result suggests that while

numerous inflammatory cytokines have been implicated in COVID-19 pathogenesis, a TNF response is unique to COVID-19-associated AHRF compared with other conditions requiring admittance to a critical care ward and provides evidence that TNF inhibitors may be a viable option for treatment specifically of COVID-19 [56]. In addition, inclusion of the PC gene *SDC1* in the final gene list distinguishing COVID-19 from non-COVID-19 ICU patients is in agreement with our previous work in which PC enrichment was the major difference in these cohorts [14].

Our innovative ML pipeline extends and improves upon previous efforts at ML-based stratification of COVID-19 patients [25–33]. A main difference in our approach is the use of whole blood gene expression data, further reduced into groups of gene modules indicative of pathologic cell populations and processes associated with SARS-CoV2 infection and COVID-19 severity. This is in contrast to ML studies utilizing patient demographic/medical record data or lung CT scan images which are limited by the quality and quantity of input data for ML algorithms [25–28]. Among other ML classification studies using transcriptomic data, our approach is the first to identify the most informative genes from within gene modules previously identified to be abnormally enriched in whole blood of COVID-19 patients. The use of whole blood RNA-seq samples rather than gene expression from sorted single-cell populations [32], peripheral blood mononuclear cells (PBMCs) [32], or throat swabs [31] provided the basis of a clinically useful blood test in which sample handling was minimal. Moreover, this work represents the first study to build transcriptome-based ML models predictive of variations in COVID-19 severity as well as differences from other individuals in the ICU, which were previously assessed separately [29,33].

Overall, our iterative ML pipeline for COVID-19 patient stratification offers a robust method of identifying critical inflammatory processes driving progression to severe disease. In addition, the excellent performance of the resulting algorithms instills confidence that they can be applied to predict disease trajectories of newly diagnosed individuals as demonstrated by the effectiveness in distinguishing severe COVID-19 patients admitted to the ICU from non-ICU COVID-19 patients. We propose that this approach be used for development of rapidly deployable blood-based PCR tests employing a small number of highly informative genes to provide clinical support for COVID-19 patient risk assessment in the clinic.

4. Materials and Methods

4.1. Study Datasets

Three publicly available whole blood bulk RNA-seq datasets previously analyzed for characterization of COVID-19 patients with varying degrees of severity were used as input for machine learning classification algorithms and an additional dataset was analyzed for validation. Raw SRA files were downloaded from Gene Expression Omnibus (GEO) accessions: GSE161731 [19], GSE172114 [22], PRJNA777938 [14], and GSE157103 [21], converted into FASTQ files, and processed into $\log_2$ gene expression values as previously described [14].

4.2. Gene Set Variation Analysis (GSVA)

The R/Bioconductor package GSVA [57] (v1.36.3) was used as a non-parametric, unsupervised method for determining enrichment of pre-defined gene modules in individual gene expression samples as previously described [14]. GSVA enrichment scores for 40 gene modules (Table S1) previously used for characterization of immune profiles in blood from COVID-19 patients [13,14] were calculated for each patient and used as input for the iterative machine learning pipeline.

4.3. Iterative Machine Learning (ML) Pipeline

The iterative ML pipeline is outlined in Figure 1. Briefly, GSVA enrichment scores for 40 gene modules (Table S1) from publicly available whole blood bulk RNA-seq datasets were used as input for 9 ML algorithms in each of 4 binary classifications. The pipeline

consisted of 3 ML iterations each of which was repeated 10 times. After the first 2 ML iterations, the top 50% of features from the top 5 performing algorithms were used as input for the next iteration. Then, the top 10 gene modules were filtered to remove redundant genes with correlation coefficients >0.8 and $\log_2$ gene expression values from the remaining genes were normalized across datasets. Normalized z-scores of gene expression values were used as input for the final ML iteration to select the top 20 gene features for each classifier. Detailed methods for each step in the pipeline are described below and available through the following link: https://github.com/adaamen/iterative-ML-COVID-pipeline (accessed on 2 February 2023).

4.4. ML Classification Algorithms

Binary classifications were carried out in Python (v3.8.2) using scikit-learn (v0.24.1) [58]. Nine ML algorithms were implemented to distinguish COVID-19 patients from healthy individuals, critical from non-critical COVID-19 patients, and COVID-19 patients admitted to the ICU from non-COVID ICU patients (logistic regression (LR), random forest (RF), support vector machine (SVM), decision tree (DTREE), AdaBoost (ADB), Gaussian naïve Bayes (NB), linear discriminant analysis (LDA), k-nearest neighbors (KNN), and gradient boosting classifier (GB)). Synthetic minority oversampling technique (SMOTE) [59] was used to account for class imbalances between patient groups used for each classification by expanding the number of samples in the minority class to ensure equal numbers between classes used for ML. For each classification, datasets were split into 70% training and 30% validation and SMOTE was applied to the training data. Classification models for each ML algorithm were then built on the training set using default parameters from the scikit-learn library and evaluated on the validation set. Average algorithm performance was calculated based on the following metrics: sensitivity, specificity, accuracy, and areas under the receiver operating characteristic (ROC) and precision-recall (PR) curves. ROC and PR curves were plotted using the Python library matplotlib (v3.3.4) [60]. Binary classification of ICU and non-ICU COVID patients from an independent dataset were used as validation for results of the iterative ML pipeline. For this classification, SMOTE was not applied. To optimize model performance on the validation dataset, a 75% training, 25% testing split was used and the class_weight = 'balanced' parameter was added to the SVM algorithm.

4.5. Feature Importance Calculation

Feature importance was calculated for the top 5 performing algorithms in each ML iteration. For each input feature a higher feature importance score indicates how relevant that feature is towards the output of each classifier. The method of calculating feature importance was determined based the appropriate metric for each algorithm and implemented through the scikit-learn Python library. Gini importance, or average decrease in impurity, was used for RF, DTREE, ADB, GB, LR, and LDA, while permutation importance was used for SVM, KNN, and NB.

4.6. Gene Feature Correlation Analysis

Feature redundancy was calculated for genes in the top 10 modules of each classifier using Pearson correlations between each feature and every other feature. When two genes shared a correlation coefficient >0.8, the gene with the highest degree of correlation with all other genes in the module was removed. Pearson correlations were computed using the cor() function in R and plotted using the corplot library.

4.7. z-Score Gene Expression Normalization

Log_2 transformed gene expression values were normalized before the final ML iteration to account for batch effects across separate datasets. To do this, z-scores were calculated for each gene across samples in each dataset using the scale() function in R and then normalized datasets were combined into a dataframe used as input for the 9 ML clas-

sification algorithms. The pre- and post-normalization expression values for each dataset combination are summarized as boxplots (Figures S1–S4A).

4.8. SHAP Value Calculation

The precise contribution (magnitude and directionality) of the top 20 gene features output from each classifier was determined using Shapley additive explanations (SHAP) [61]. SHAP values were calculated based on the RF algorithm for the final iteration of each classifier. SHAP summary plots were visualized in Python using the shap library (v0.39.0).

4.9. Statistical Analysis

Normalized gene expression values of the top 20 genes were compared between the two classes of each binary classification using multiple unpaired t-tests with Welch's correction in GraphPad Prism (v9.1.0; www.graphpad.com, (accessed on 20 December 2022)). A two-tailed p-value < 0.05 was considered statistically significant.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24054905/s1.

Author Contributions: Conceptualization, A.R.D., A.C.G. and P.E.L.; methodology, A.R.D. and P.B.; software, A.R.D. and P.B.; validation, A.R.D. and P.B.; formal analysis, A.R.D.; investigation, A.R.D.; resources, P.B., A.C.G. and P.E.L.; data curation, P.B.; writing—original draft preparation, A.R.D.; writing—review and editing, A.R.D. and P.E.L.; visualization, A.R.D.; supervision, A.C.G. and P.E.L.; project administration, A.C.G. and P.E.L.; funding acquisition, A.C.G. and P.E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the RILITE Foundation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The RNA-seq data analyzed in the current study are publicly available through the National Center for Biotechnology Information (NCBI) under Sequence Read Archive (SRA) project number PRJNA777938 and the Gene Expression Omnibus (GEO) under accessions GSE161731 and GSE172114.

Acknowledgments: We thank the patients who provided samples analyzed for this study as well as the authors who made their data publicly available and facilitated our analyses.

Conflicts of Interest: The authors declare no conflict of interest. There is no conflict of interest with this work and AMPEL BioSolutions, LLC.

References

1. Guan, W.-J.; Ni, Z.-Y.; Hu, Y.; Liang, W.-H.; Ou, C.-Q.; He, J.-X.; Liu, L.; Shan, H.; Lei, C.-L.; Hui, D.S.C.; et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **2020**, *382*, 1708–1720. [CrossRef] [PubMed]
2. Wang, D.; Hu, B.; Hu, C.; Zhu, F.; Liu, X.; Zhang, J.; Wang, B.; Xiang, H.; Cheng, Z.; Xiong, Y.; et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA J. Am. Med. Assoc.* **2020**, *323*, 1061–1069. [CrossRef] [PubMed]
3. WHO. WHO Coronavirus WHO Coronavirus. Available online: https://covid19.who.int/ (accessed on 1 January 2023).
4. Wiersinga, W.J.; Rhodes, A.; Cheng, A.C.; Peacock, S.J.; Prescott, H.C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* **2020**, *324*, 782–793. [CrossRef] [PubMed]
5. Nalbandian, A.; Sehgal, K.; Gupta, A.; Madhavan, M.V.; McGroder, C.; Stevens, J.S.; Cook, J.R.; Nordvig, A.S.; Shalev, D.; Sehrawat, T.S.; et al. Post-Acute COVID-19 Syndrome. *Nat. Med.* **2021**, *27*, 601–615. [CrossRef]
6. Williamson, E.J.; Walker, A.J.; Bhaskaran, K.; Bacon, S.; Bates, C.; Morton, C.E.; Curtis, H.J.; Mehrkar, A.; Evans, D.; Inglesby, P.; et al. Factors Associated with COVID-19-Related Death Using OpenSAFELY. *Nature* **2020**, *584*, 430–436. [CrossRef]
7. Takahashi, T.; Ellingson, M.K.; Wong, P.; Israelow, B.; Lucas, C.; Klein, J.; Silva, J.; Mao, T.; Oh, J.E.; Tokuyama, M.; et al. Sex Differences in Immune Responses That Underlie COVID-19 Disease Outcomes. *Nature* **2020**, *588*, 315–320. [CrossRef]
8. Mokhtari, T.; Hassani, F.; Ghaffari, N.; Ebrahimi, B.; Yarahmadi, A.; Hassanzadeh, G. COVID-19 and Multiorgan Failure: A Narrative Review on Potential Mechanisms. *J. Mol. Histol.* **2020**, *51*, 613–628. [CrossRef]
9. Michalski, J.E.; Kurche, J.S.; Schwartz, D.A. From ARDS to Pulmonary Fibrosis: The next Phase of the COVID-19 Pandemic? *Transl. Res.* **2022**, *241*, 13–24. [CrossRef]

10. Chen, Y.-T.; Shao, S.-C.; Hsu, C.-K.; Wu, I.-W.; Hung, M.-J.; Chen, Y.-C. Incidence of Acute Kidney Injury in COVID-19 Infection: A Systematic Review and Meta-Analysis. *Crit. Care* **2020**, *24*, 346. [CrossRef]
11. Merad, M.; Blish, C.A.; Sallusto, F.; Iwasaki, A. The Immunology and Immunopathology of COVID-19. *Science* **2022**, *375*, 1122–1127. [CrossRef]
12. Ilieva, M.; Tschaikowski, M.; Vandin, A.; Uchida, S. The Current Status of Gene Expression Profilings in COVID-19 Patients. *Clin. Transl. Discov.* **2022**, *2*, e104. [CrossRef]
13. Daamen, A.R.; Bachali, P.; Owen, K.A.; Kingsmore, K.M.; Hubbard, E.L.; Labonte, A.C.; Robl, R.; Shrotri, S.; Grammer, A.C.; Lipsky, P.E. Comprehensive Transcriptomic Analysis of COVID-19 Blood, Lung, and Airway. *Sci. Rep.* **2021**, *11*, 7052. [CrossRef]
14. Daamen, A.R.; Bachali, P.; Bonham, C.A.; Somerville, L.; Sturek, J.M.; Grammer, A.C.; Kadl, A.; Lipsky, P.E. COVID-19 Patients Exhibit Unique Transcriptional Signatures Indicative of Disease Severity. *Front. Immunol.* **2022**, *13*, 989556. [CrossRef]
15. Mathew, D.; Giles, J.R.; Baxter, A.E.; Oldridge, D.A.; Greenplate, A.R.; Wu, J.E.; Alanio, C.; Kuri-Cervantes, L.; Pampena, M.B.; D'Andrea, K.; et al. Deep Immune Profiling of COVID-19 Patients Reveals Distinct Immunotypes with Therapeutic Implications. *Science* **2020**, *369*, eabc8511. [CrossRef]
16. Lucas, C.; Wong, P.; Klein, J.; Castro, T.B.R.; Silva, J.; Sundaram, M.; Ellingson, M.K.; Mao, T.; Oh, J.E.; Israelow, B.; et al. Longitudinal Analyses Reveal Immunological Misfiring in Severe COVID-19. *Nature* **2020**, *584*, 463–469. [CrossRef]
17. Wilk, A.J.; Rustagi, A.; Zhao, N.Q.; Roque, J.; Martínez-Colón, G.J.; McKechnie, J.L.; Ivison, G.T.; Ranganath, T.; Vergara, R.; Hollis, T.; et al. A Single-Cell Atlas of the Peripheral Immune Response in Patients with Severe COVID-19. *Nat. Med.* **2020**, *26*, 1070–1076. [CrossRef]
18. Wilk, A.J.; Lee, M.J.; Wei, B.; Parks, B.; Pi, R.; Martínez-Colón, G.J.; Ranganath, T.; Zhao, N.Q.; Taylor, S.; Becker, W.; et al. Multi-Omic Profiling Reveals Widespread Dysregulation of Innate Immunity and Hematopoiesis in COVID-19. *J. Exp. Med.* **2021**, *218*, e20210582. [CrossRef]
19. McClain, M.T.; Constantine, F.J.; Henao, R.; Liu, Y.; Tsalik, E.L.; Burke, T.W.; Steinbrink, J.M.; Petzold, E.; Nicholson, B.P.; Rolfe, R.; et al. Dysregulated Transcriptional Responses to SARS-CoV-2 in the Periphery. *Nat. Commun.* **2021**, *12*, 1079. [CrossRef]
20. Stephenson, E.; Reynolds, G.; Botting, R.A.; Calero-Nieto, F.J.; Morgan, M.D.; Tuong, Z.K.; Bach, K.; Sungnak, W.; Worlock, K.B.; Yoshida, M.; et al. Single-Cell Multi-Omics Analysis of the Immune Response in COVID-19. *Nat. Med.* **2021**, *27*, 904–916. [CrossRef]
21. Overmyer, K.A.; Shishkova, E.; Miller, I.J.; Balnis, J.; Bernstein, M.N.; Peters-Clarke, T.M.; Meyer, J.G.; Quan, Q.; Muehlbauer, L.K.; Trujillo, E.A.; et al. Large-Scale Multi-Omic Analysis of COVID-19 Severity. *Cell Syst.* **2021**, *12*, 23–40.e7. [CrossRef]
22. Carapito, R.; Li, R.; Helms, J.; Carapito, C.; Gujja, S.; Rolli, V.; Guimaraes, R.; Malagon-Lopez, J.; Spinnhirny, P.; Lederle, A.; et al. Identification of Driver Genes for Critical Forms of COVID-19 in a Deeply Phenotyped Young Patient Cohort. *Sci. Transl. Med.* **2022**, *14*, eabj7521. [CrossRef]
23. Shah, P.; Kendall, F.; Khozin, S.; Goosen, R.; Hu, J.; Laramie, J.; Ringel, M.; Schork, N. Artificial Intelligence and Machine Learning in Clinical Development: A Translational Perspective. *NPJ Digit. Med.* **2019**, *2*, 69. [CrossRef]
24. Meraihi, Y.; Gabis, A.B.; Mirjalili, S.; Ramdane-Cherif, A.; Alsaadi, F.E. Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. *SN Comput. Sci.* **2022**, *3*, 286. [CrossRef] [PubMed]
25. Quiroz-Juárez, M.A.; Torres-Gómez, A.; Hoyo-Ulloa, I.; León-Montiel, R.d.J.; U'Ren, A.B. Identification of High-Risk COVID-19 Patients Using Machine Learning. *PLoS ONE* **2021**, *16*, e0257234. [CrossRef] [PubMed]
26. Guhan, B.; Almutairi, L.; Sowmiya, S.; Snekhalatha, U.; Rajalakshmi, T.; Aslam, S.M. Automated System for Classification of COVID-19 Infection from Lung CT Images Based on Machine Learning and Deep Learning Techniques. *Sci. Rep.* **2022**, *12*, 17417. [CrossRef] [PubMed]
27. Nguyen, D.; Kay, F.; Tan, J.; Yan, Y.; Ng, Y.S.; Iyengar, P.; Peshock, R.; Jiang, S. Deep Learning–Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability. *Front. Artif. Intell.* **2021**, *4*, 694875. [CrossRef]
28. Zargari Khuzani, A.; Heidari, M.; Shariati, S.A. COVID-Classifier: An Automated Machine Learning Model to Assist in the Diagnosis of COVID-19 Infection in Chest X-Ray Images. *Sci. Rep.* **2021**, *11*, 9887. [CrossRef]
29. Penrice-Randal, R.; Dong, X.; Shapanis, A.G.; Gardner, A.; Harding, N.; Legebeke, J.; Lord, J.; Vallejo, A.F.; Poole, S.; Brendish, N.J.; et al. Blood Gene Expression Predicts Intensive Care Unit Admission in Hospitalised Patients with COVID-19. *Front. Immunol.* **2022**, *13*, 988685. [CrossRef]
30. Zarei Ghobadi, M.; Emamzadeh, R.; Teymoori-Rad, M.; Afsaneh, E. Exploration of Blood-Derived Coding and Non-Coding RNA Diagnostic Immunological Panels for COVID-19 through a Co-Expressed-Based Machine Learning Procedure. *Front. Immunol.* **2022**, *13*, 1001070. [CrossRef]
31. Song, X.; Zhu, J.; Tan, X.; Yu, W.; Wang, Q.; Shen, D.; Chen, W. XGBoost-Based Feature Learning Method for Mining COVID-19 Novel Diagnostic Markers. *Front. Public Health* **2022**, *10*, 926069. [CrossRef]
32. Li, H.; Huang, F.; Liao, H.; Li, Z.; Feng, K.; Huang, T.; Cai, Y.-D. Identification of COVID-19-Specific Immune Markers Using a Machine Learning Method. *Front. Mol. Biosci.* **2022**, *9*, 952626. [CrossRef] [PubMed]
33. Li, X.; Zhou, X.; Ding, S.; Chen, L.; Feng, K.; Li, H.; Huang, T.; Cai, Y.-D. Identification of Transcriptome Biomarkers for Severe COVID-19 with Machine Learning Methods. *Biomolecules* **2022**, *12*, 1735. [CrossRef] [PubMed]
34. Lohmann, S.; Herold, A.; Bergauer, T.; Belousov, A.; Betzl, G.; Demario, M.; Dietrich, M.; Luistro, L.; Poignée-Heger, M.; Schostack, K.; et al. Gene Expression Analysis in Biomarker Research and Early Drug Development Using Function Tested Reverse Transcription Quantitative Real-Time PCR Assays. *Methods* **2013**, *59*, 10–19. [CrossRef] [PubMed]

35. El-Deiry, W.S.; Goldberg, R.M.; Lenz, H.-J.; Shields, A.F.; Gibney, G.T.; Tan, A.R.; Brown, J.; Eisenberg, B.; Heath, E.I.; Phuphanich, S.; et al. The Current State of Molecular Testing in the Treatment of Patients with Solid Tumors, 2019. CA. *Cancer J. Clin.* **2019**, *69*, 305–343. [CrossRef]
36. Marshall, K.W.; Mohr, S.; Khettabi, F.E.; Nossova, N.; Chao, S.; Bao, W.; Ma, J.; Li, X.-J.; Liew, C.-C. A Blood-Based Biomarker Panel for Stratifying Current Risk for Colorectal Cancer. *Int. J. Cancer* **2010**, *126*, 1177–1186. [CrossRef] [PubMed]
37. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281. [CrossRef]
38. Mazlan, A.U.; Sahabudin, N.A.; Remli, M.A.; Ismail, N.S.; Mohamad, M.S.; Nies, H.W.; Abd Warif, N.B. A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. *Processes* **2021**, *9*, 1466. [CrossRef]
39. Kingsmore, K.M.; Puglisi, C.E.; Grammer, A.C.; Lipsky, P.E. An Introduction to Machine Learning and Analysis of Its Use in Rheumatic Diseases. *Nat. Rev. Rheumatol.* **2021**, *17*, 710–730. [CrossRef]
40. Saeys, Y.; Inza, I.; Larranaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
41. Ang, J.C.; Mirzal, A.; Haron, H.; Hamed, H.N.A. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2016**, *13*, 971–989. [CrossRef]
42. Samy, A.; Maher, M.A.; Abdelsalam, N.A.; Badr, E. SARS-CoV-2 Potential Drugs, Drug Targets, and Biomarkers: A Viral-Host Interaction Network-Based Analysis. *Sci. Rep.* **2022**, *12*, 11934. [CrossRef]
43. Cavalcante, L.T.D.F.; da Fonseca, G.C.; Amado Leon, L.A.; Salvio, A.L.; Brustolini, O.J.; Gerber, A.L.; Guimarães, A.P.D.C.; Marques, C.A.B.; Fernandes, R.A.; Ramos Filho, C.H.F.; et al. Buffy Coat Transcriptomic Analysis Reveals Alterations in Host Cell Protein Synthesis and Cell Cycle in Severe COVID-19 Patients. *Int. J. Mol. Sci.* **2022**, *23*, 13588. [CrossRef] [PubMed]
44. Prado, C.A.D.S.; Fonseca, D.L.M.; Singh, Y.; Filgueiras, I.S.; Baiocchi, G.C.; Plaça, D.R.; Marques, A.H.; Dantas-Komatsu, R.C.S.; Usuda, J.N.; Freire, P.P.; et al. Integrative Systems Immunology Uncovers Molecular Networks of the Cell Cycle That Stratify COVID-19 Severity. *J. Med. Virol.* **2023**, *95*, e28450. [CrossRef] [PubMed]
45. Duan, C.; Ma, R.; Zeng, X.; Chen, B.; Hou, D.; Liu, R.; Li, X.; Liu, L.; Li, T.; Huang, H. SARS-CoV-2 Achieves Immune Escape by Destroying Mitochondrial Quality: Comprehensive Analysis of the Cellular Landscapes of Lung and Blood Specimens From Patients With COVID-19. *Front. Immunol.* **2022**, *13*, 946731. [CrossRef] [PubMed]
46. Chernyak, B.V.; Popova, E.N.; Prikhodko, A.S.; Grebenchikov, O.A.; Zinovkina, L.A.; Zinovkin, R.A. COVID-19 and Oxidative Stress. *Biochemistry* **2020**, *85*, 1543–1553. [CrossRef] [PubMed]
47. Guarnieri, J.W.; Dybas, J.M.; Fazelinia, H.; Kim, M.S.; Frere, J.; Zhang, Y.; Albrecht, Y.S.; Murdock, D.G.; Angelin, A.; Singh, L.N.; et al. Targeted Down Regulation of Core Mitochondrial Genes During SARS-CoV-2 Infection. *bioRxiv* **2022**. *preprint*. [CrossRef]
48. McKenna, E.; Wubben, R.; Isaza-Correa, J.M.; Melo, A.M.; Mhaonaigh, A.U.; Conlon, N.; O'Donnell, J.S.; Ní Cheallaigh, C.; Hurley, T.; Stevenson, N.J.; et al. Neutrophils in COVID-19: Not Innocent Bystanders. *Front. Immunol.* **2022**, *13*, 2548. [CrossRef]
49. Aschenbrenner, A.C.; Mouktaroudi, M.; Kraemer, B.; Antonakos, N.; Oestreich, M.; Gkizeli, K.; Nuesch-Germano, M.; Saridaki, M.; Bonaguro, L.; Reusch, N.; et al. Disease Severity-Specific Neutrophil Signatures in Blood Transcriptomes Stratify COVID-19 Patients. *medRxiv* **2020**, *13*, 1–25. [CrossRef]
50. Schulte-Schrepping, J.; Reusch, N.; Paclik, D.; Baßler, K.; Schlickeiser, S.; Zhang, B.; Krämer, B.; Krammer, T.; Brumhard, S.; Bonaguro, L.; et al. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **2020**, *182*, 1419–1440.e23. [CrossRef]
51. Lee, J.S.; Park, S.; Jeong, H.W.; Ahn, J.Y.; Choi, S.J.; Lee, H.; Choi, B.; Nam, S.K.; Sa, M.; Kwon, J.S.; et al. Immunophenotyping of Covid-19 and Influenza Highlights the Role of Type i Interferons in Development of Severe Covid-19. *Sci. Immunol.* **2020**, *5*, eabd1554. [CrossRef]
52. Dong, Z.; Yan, Q.; Cao, W.; Liu, Z.; Wang, X. Identification of Key Molecules in COVID-19 Patients Significantly Correlated with Clinical Outcomes by Analyzing Transcriptomic Data. *Front. Immunol.* **2022**, *13*, 930866. [CrossRef] [PubMed]
53. Zhou, R.; To, K.K.-W.; Wong, Y.-C.; Liu, L.; Zhou, B.; Li, X.; Huang, H.; Mo, Y.; Luk, T.-Y.; Lau, T.T.-K.; et al. Acute SARS-CoV-2 Infection Impairs Dendritic Cell and T Cell Responses. *Immunity* **2020**, *53*, 864–877.e5. [CrossRef] [PubMed]
54. Chen, Z.; John Wherry, E. T Cell Responses in Patients with COVID-19. *Nat. Rev. Immunol.* **2020**, *20*, 529–536. [CrossRef] [PubMed]
55. Cross, A.R.; de Andrea, C.E.; Villalba-Esparza, M.; Landecho, M.F.; Cerundolo, L.; Weeratunga, P.; Etherington, R.E.; Denney, L.; Ogg, G.; Ho, L.-P.; et al. Spatial Transcriptomic Characterization of COVID-19 Pneumonitis Identifies Immune Circuits Related to Tissue Injury. *JCI Insight* **2022**, *8*, e157837. [CrossRef]
56. Izadi, Z.; Brenner, E.J.; Mahil, S.K.; Dand, N.; Yiu, Z.Z.N.; Yates, M.; Ungaro, R.C.; Zhang, X.; Agrawal, M.; Colombel, J.-F.; et al. Association Between Tumor Necrosis Factor Inhibitors and the Risk of Hospitalization or Death Among Patients with Immune-Mediated Inflammatory Disease and COVID-19. *JAMA Netw. Open* **2021**, *4*, e2129639. [CrossRef]
57. Hänzelmann, S.; Castelo, R.; Guinney, J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinform.* **2013**, *14*, 7. [CrossRef] [PubMed]
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
59. Blagus, R.; Lusa, L. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinform.* **2013**, *14*, 106. [CrossRef]

60. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
61. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U., Von Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.