



Article

Integrating AI/ML Models for Patient Stratification Leveraging Omics Dataset and Clinical Biomarkers from COVID-19 Patients: A Promising Approach to Personalized Medicine

Babatunde Bello [†], Yogesh N. Bunday [†], Roshan Bhawe, Maksim Khotimchenko , Szczepan W. Baran , Kaushik Chakravarty * and Jyotika Varshney *

VeriSIM Life Inc., 1 Sansome Street, Suite 3500, San Francisco, CA 94104, USA

* Correspondence: kaushik.chakravarty@verisimlife.com (K.C.); jo.varshney@verisimlife.com (J.V.)

[†] These authors contributed equally to this work.

Abstract: The COVID-19 pandemic has presented an unprecedented challenge to the healthcare system. Identifying the genomics and clinical biomarkers for effective patient stratification and management is critical to controlling the spread of the disease. Omics datasets provide a wealth of information that can aid in understanding the underlying molecular mechanisms of COVID-19 and identifying potential biomarkers for patient stratification. Artificial intelligence (AI) and machine learning (ML) algorithms have been increasingly used to analyze large-scale omics and clinical datasets for patient stratification. In this manuscript, we demonstrate the recent advances and predictive accuracies in AI- and ML-based patient stratification modeling linking omics and clinical biomarker datasets, focusing on COVID-19 patients. Our ML model not only demonstrates that clinical features are enough of an indicator of COVID-19 severity and survival, but also infers what clinical features are more impactful, which makes our approach a useful guide for clinicians for prioritization best-fit therapeutics for a given cohort of patients. Moreover, with weighted gene network analysis, we are able to provide insights into gene networks that have a significant association with COVID-19 severity and clinical features. Finally, we have demonstrated the importance of clinical biomarkers in identifying high-risk patients and predicting disease progression.

Keywords: SARS-CoV-2; COVID-19; artificial intelligence; omics; patient stratification; risk management



Citation: Bello, B.; Bunday, Y.N.; Bhawe, R.; Khotimchenko, M.; Baran, S.W.; Chakravarty, K.; Varshney, J. Integrating AI/ML Models for Patient Stratification Leveraging Omics Dataset and Clinical Biomarkers from COVID-19 Patients: A Promising Approach to Personalized Medicine. *Int. J. Mol. Sci.* **2023**, *24*, 6250. <https://doi.org/10.3390/ijms24076250>

Academic Editor: Jordi Pérez-Tur

Received: 28 February 2023

Revised: 17 March 2023

Accepted: 24 March 2023

Published: 26 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

Over the past two years, the global COVID-19 pandemic has highlighted the crucial role of accurate patient diagnostics in preventing an overload of healthcare resources. The pandemic has led to a significant increase in patient surges globally, resulting in varying degrees of respiratory illnesses and a rise in mortality rates [1]. The symptoms of COVID-19 induced by the varying strains of the pathogen SARS-CoV-2 are challenging to differentiate from other common respiratory infections in a large proportion of those infected. Moreover, disease progression varies from asymptomatic cases to critical patient conditions, making it challenging for appropriate treatment selections and accurate prognoses. Some patients with COVID-19 rapidly develop severe dysfunctions and even critical illness, demonstrating an excellent example of infection variability due to internal physiological factors [2]. Various SARS-CoV-2 variants led to more concerns due to lowered vaccine efficacy and rapidly decreasing immunity. On the other hand, waves and variants of SARS-CoV-2 helped us understand the causes and effects of the COVID-19 catastrophe that was related to respiratory illness and multiorgan failure. The emergence of novel omicron hybrid variants such as BA (x), XE, XD, XF, etc., demonstrated the role of population variability in disease severity and outcome [3]. Furthermore, even after recovery, a portion of patients keep experiencing a spectrum of COVID-19-like symptoms generally termed “Long COVID”, making disease outcome prognosis even more challenging. Subsequently, this

situation is substantially aggravated by the lack of understanding of the main signaling and pathogenetic mechanisms induced by COVID-19 infection in the host.

Correct risk evaluation and management in the realm of infectious diseases are critical for the diagnostic assessment and selection of the best line of therapies, markedly increasing the likelihood of patient survival [4]. The COVID-19 outbreak has highlighted the significance of implementing these measures. Therefore, one of the greatest challenges during the pandemic, especially in resource-strained settings, has been the early identification of individual patients at higher risk for adverse outcomes. Hence, it is critical to develop and implement intelligent risk-assessment tools that can predict a patient's disease progression and recovery and suggest best-fit therapeutics for markedly reducing disease severity.

Data-driven risk evaluation based on artificial intelligence/machine learning (AI/ML) software models is considered an effective decision-making algorithm in the streamlining triage of emergency patients [5]. Sorting patients into different cohorts based on the disease severity and prognosis helps rationally allocate healthcare resources and identify the most effective therapy. AI-powered models have been recently used with success to simulate and predict drug–drug interactions in different patient groups, which aids in preventing potential issues during the initial stages of drug development in clinical trials [6]. Contemporary AI/ML-driven approaches in the healthcare space have accelerated the drug discovery and development pipelines due to the availability and prioritization of data alongside a wide range of statistical learning methods that leverage data to draw inferences and make new predictions [7]. AI technologies are able to recognize specific pathways responsible for the development of disease-related syndromes that leads to a more accurate categorization of the patient cohorts [8]. Additionally, computational assistance to healthcare providers contributes to reduced emotional pressure and, consequently, increased accuracy selection of successful therapeutic strategies [9,10].

Implementation of AI/ML-based patient stratification is a suitable strategy for developing custom patient-specific therapeutic guidance because of the sufficient availability of COVID-19 clinical data. Such strategies may contribute to overcoming the pandemic altogether as model applications move from being patient-based to population-based [5]. Current AI/ML algorithms providing COVID-19 prognosis belong to either image- or non-image-based categories. The first group significantly relies on chest X-ray or digital tomography readouts [11] that reflect lung function but cannot accurately quantitate the degree of immune response and upregulation of biomarkers in the inflammatory pathways. It was demonstrated that quantifying proinflammatory cytokine levels, blood, and urine biomarkers improves the assessments of patient clinical conditions, showing high accuracy [12]. Non-image-based methods for COVID-19 include using electronic health record (EHR) information to diagnose COVID-19 and assess its severity. These methods consist of two basic types of studies: score system-based and ML-based COVID-19 diagnostics [13]. Researchers seek to identify important predictors and assign them associated scores for the former. Subsequently, the summation of the scores can potentially lead to the stratification of patients with different disease severities [12,14]. Currently, several risk assessment scores are available to predict the severity of different, none-COVID-19-related diseases for ICU patients [15]. Furthermore, predictors indicating the need for intensive respiratory or vasopressor support for COVID-19 patients or suggesting the high mortality risk in COVID-19 patients with pneumonia have been identified [16,17].

Liang et al. [12] built a predictive risk score (COVID-GRAM) system, which included ten important predictive factors screened from 72 potential predictors among epidemiological, clinical, laboratory, and imaging variables. The COVID-GRAM was used to estimate the risk of developing critical illness for patients with COVID-19. Other methods are based on a gradient boosting decision tree model that predicts an individual's COVID-19 infection progression, recovery rate, and mortality risk using demographic information such as age, sex, race, and a series of routine lab tests [18]. Recent advances in deep learning (DL) combined with EHR datasets have gained recent traction for diagnosing COVID-19. For example, a feedforward neural network-based DL survival model was used to predict

critical illness development risk in COVID-19 patients by evaluating 74 baseline clinical features [12]. It has been noted that patient genomics data, which play a crucial role in the diversity of disease severity, are not accounted for in any of those models [19,20]. Adding genomics information into a predictive modeling system can help enrich and identify critical pathways and associated nuances contributing to the disease severity and select the most appropriate treatment options.

Understanding the biological pathways that impact severe cases of COVID-19 is crucial to identifying effective drug interventions and improving survival rates. The main objective of the study was to understand the most impactful biomarkers that contribute to severe cases and lower survival through ML predictive models (Figure 1). Models were trained on clinical patient data that include several biomarker levels in correspondence with case severity and survival. After achieving optimal performance, an analysis of model explainability was conducted to reveal the “black box effect” of the predictive models and identify the biomarkers and their corresponding values that have the greatest influence on model predictions. Upon identifying significant biomarkers, a correlation analysis was conducted between groups of genes and the biomarkers that were modulated to pinpoint the most probable cluster of genes contributing to severe cases of COVID-19 and decreased survival rates. ML modeling was employed to conduct a correlation analysis on a patient pool consisting of both OMICS and clinical datasets. Our AI/ML modeling has helped us reveal the significant association between the most influential gene cluster and the related biological pathways. This manuscript showcases our findings, which can greatly improve our comprehension of the features of potential drug candidates to address the severity and mortality of COVID-19.

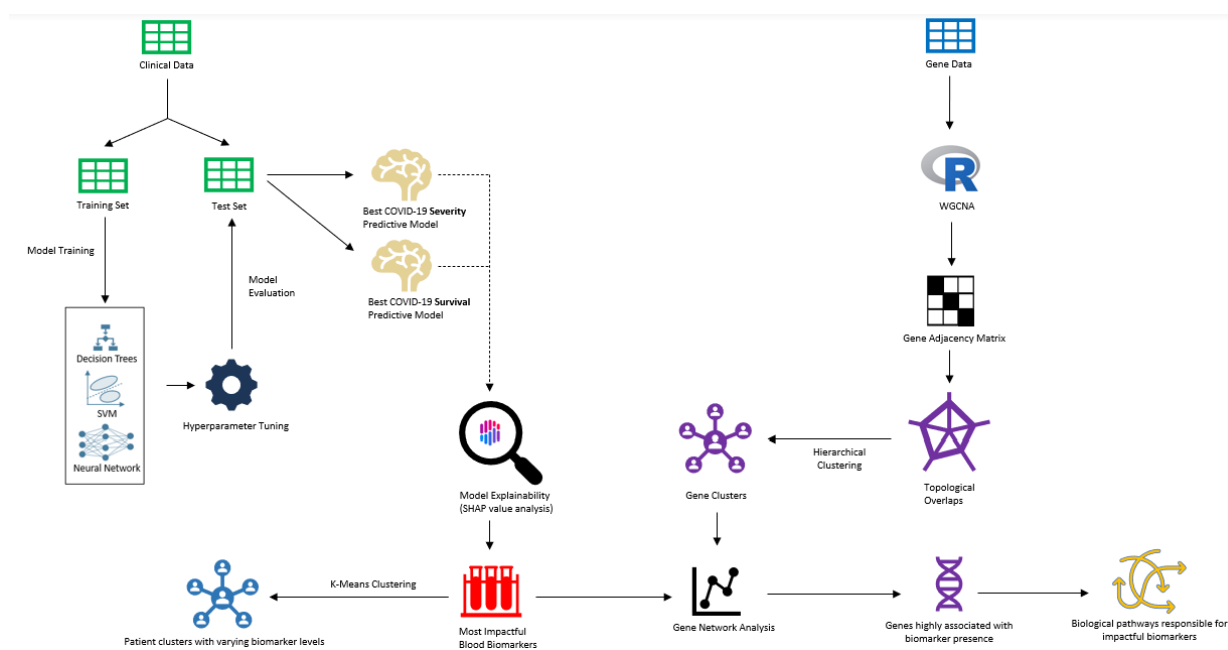


Figure 1. Overview of the workflow for an AI/ML-driven model that predicts the outcome of COVID-19 infection in patients, utilizing a set of clinical biomarkers and genomics dataset.

2. Results

2.1. Model Outcomes

Performance metrics for various modeling architectures were collected in Supplementary (Tables S6 and S7). Missing values within the dataset was imputed using Multiple Iteration Chain Estimation (MICE) regression techniques [21], where each missing data point is iteratively learned from other available features. Boosted Decision Tree architectures, such as LightGBM, XGBoost, and CatBoost, have built in techniques for handling

missing values [22], so evaluation metrics were reported twice (with or without MICE) for each of these architectures.

The balanced accuracy was 91.6%, and the ROC-AUC score was 98.1% for the best performing severity model. The balanced accuracy was 99.4%, and the ROC-AUC score was 99.9% with the best performing survival model. Both models were LightGBM classifier models, a lightweight modeling infrastructure that applies Gradient Boosting Machine (GBM) decision trees. GBMs consist of iterative, or “boosted”, decision trees that fit training feature thresholds to approximate predictions. The LightGBM model architecture incorporates missing value-handling logic that minimizes loss on the training dataset in the most effective way possible. The internal decision trees within the model itself represent missing values in such a way that it aligns most with the training data; however, it maintains the assumption that the value is unknown rather than inferring an actual value. The modeling logic’s robustness enables it to handle missing data as input, indicating that clinicians are not required to input all biomarker levels observed in the clinical training dataset columns to receive a valid prediction from the trained model. The high performing results indicate the usefulness of clinical features for both predictive use cases. The predictive modeling results are summarized in Table 1. ROC curves for severity and survival models are shown in Figures 2 and 3, respectively.

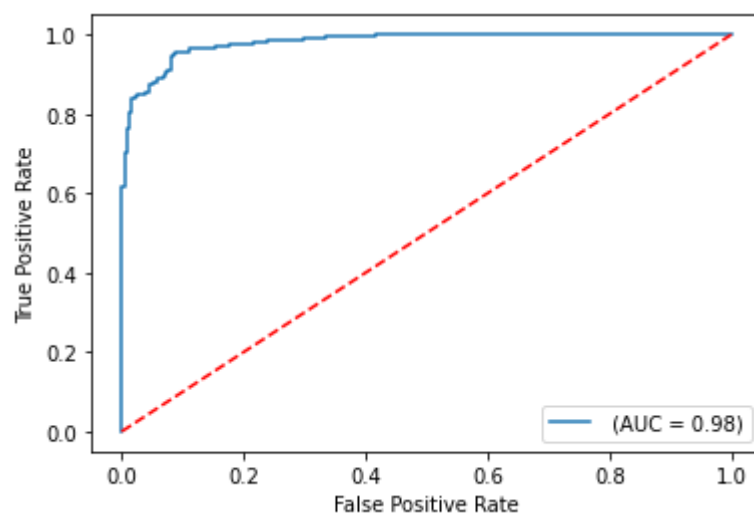


Figure 2. ROC curve on test set for COVID-19 severity prediction model.

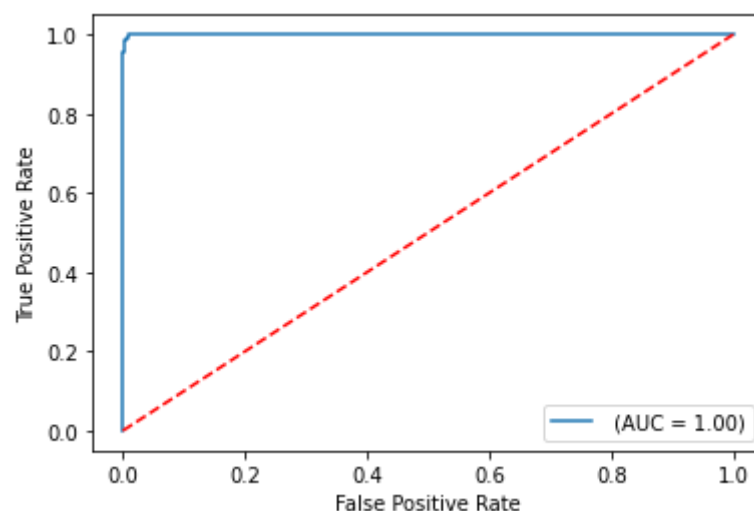


Figure 3. ROC curve on test set for COVID-19 survival prediction model.

Table 1. Evaluation metrics for COVID-19 severity and survival models.

	Severity	Survival
Balanced accuracy	91.6%	99.1%
ROC-AUC	98.1%	99.9%

The fact that the clinical features used to train both models can be generalized, along with the absence of some portions of the biomarker laboratory datasets, reinforces the practicality of deploying predictive model in a clinical setting. The input to the predictive models need not include all the listed blood biomarker values in the feature set to generate a dependable prediction, but rather only some of them.

In the next use case, training data was also modified only to include patients with no comorbidities to assess if similar biomarker-based features for non-comorbidity patients are representative of predicting COVID-19 case severity and survival for a cohort containing both comorbidity and non-comorbidity patients within the same test set. With this data filtering, the training set was reduced by 63%, making it reasonable for the evaluation metrics to drop. The results of this training data modification are shown in Table 2. ROC curves for severity and survival models are shown in Figures 4 and 5, respectively.

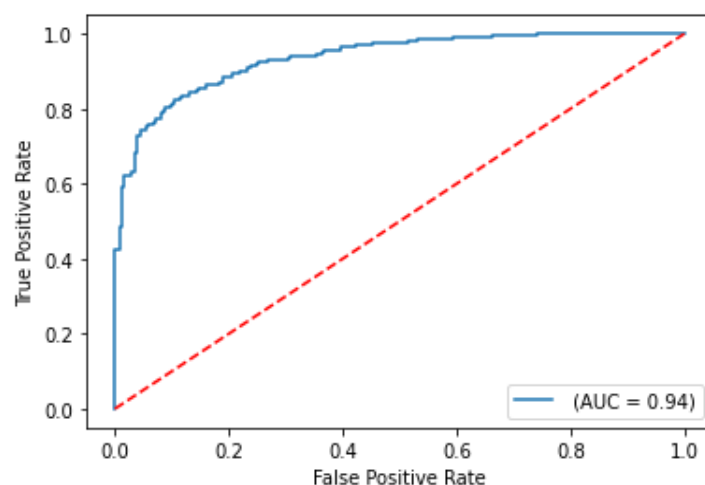
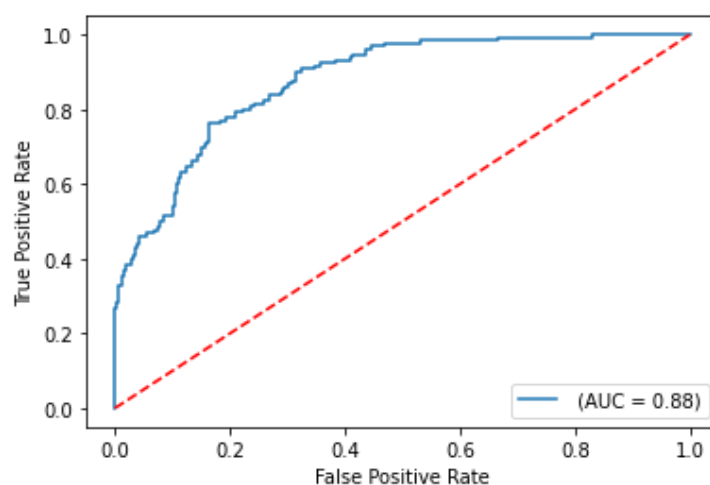
**Figure 4.** ROC curve on test set for COVID-19 severity prediction model with comorbidities re-moved from training set.**Figure 5.** ROC curve on the test set for COVID-19 survival prediction model with comorbidities removed from the training set.

Table 2. Evaluation metrics for COVID-19 severity and survival models without comorbidities in the training data.

	Severity	Survival
Balanced accuracy	85.4%	69.8%
ROC-AUC	93.5%	87.8%

Despite using a modified training set, the model performance results still demonstrate a robust test ROC-AUC score of 93.5% for predicting the severity of COVID-19. This suggests that the features without comorbidities remain highly indicative of COVID-19 case severity, regardless of whether the patients have comorbidities or not. However, the ROC-AUC score for the COVID-19 survival models is significantly lower with the training set adjustment, which indicates that certain comorbidities are strongly correlated with survival. These results suggest that it is unlikely to observe a strong influence of comorbidity features through the model explainability analysis for the COVID-19 severity model. However, impactful comorbidity features such as coronary artery disease and diabetes from model explainability analysis for the COVID-19 survival model can be noted during analysis.

2.2. Patient Biomarker Analysis

2.2.1. Clinical Biomarker Evaluation

Model explainability has transformed the way ML can be applied, especially in the biomedical domain. Through model explainability methods, inferences can be gathered on whether higher or lower feature values contribute to a higher probability of a given predictive class. As high performance for both COVID-19 severity and survival models were achieved, model explainability analysis allowed for the evaluation of the biomarker values that are highly correlated with more severe COVID-19 and least survival of COVID-19 cases. Model explainability was achieved through SHAP (SHapley Additive exPlanations) analysis, which is an approach to explain the output of any machine learning model based on game theory that infers model interpretation of specific feature values based on the impact that they have on the model itself [23]. Specific feature values are transformed into SHAP values to measure the impact on the model prediction. Equation (1) below shows the calculation of the SHAP value for the n th feature among N feature subsets, given a feature subset without the n th feature (S), the total number of features (F), and the predictive model (M):

$$SHAP_{n,M} = \sum_{S \subseteq N \setminus \{n\}} \frac{|S|!(F - \text{length}(S) - 1)!}{F!} [M(S \cup \{n\}) - M(S)] \quad (1)$$

The feature impact scores interpolated from the SHAP values for the top 20 most impactful features of the COVID-19 severity and survival predictive models are shown in Figures 6 and 7, where higher feature impact scores indicated whether higher or lower feature values were correlated with more severe COVID-19 cases and lower chance of survival from COVID-19. The top 20 features were determined by the highest absolute SHAP value mean, which shows the most impactful features contributing to either class (severe case/moderate case) in Figure 6 (survival/mortality) and in Figure 7.

As implied by the results presented earlier, there were no significant comorbidity-based features in the COVID-19 severity model. However, it was interesting to note that the presence of comorbidities, coronary artery disease, and diabetes were strongly correlated with low survival. Moreover, there has been evidence to suggest the presence of coronary artery disease [24] and diabetes [25] as a predictor for both COVID-19 severity and mortality. It is also interesting to note that patients who had been given methylprednisolone had a higher probability of having a severe case and lower survivability. Although there is no established link between the identified medication and COVID-19 severity or patient survival in the clinical dataset, it would be valuable to examine the medication's impact when used in conjunction with insights from impactful biomarkers. Furthermore, this is

particularly relevant given that steroid anti-inflammatory drugs are frequently given to patients with severe and critical respiratory infections [26]. It is highly possible that severe patients were given the medication as a potentially preventive measure; however, from the analysis there is no indication of any positive or negative effect on their survival or severity.

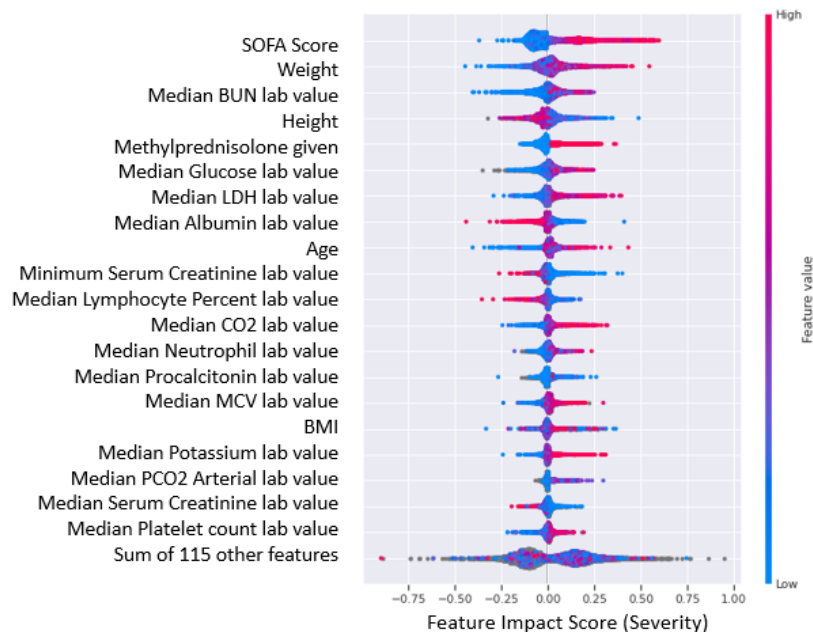


Figure 6. Top 20 most impactful features for COVID-19 severity predictive model. The red data points indicate higher values for the particular feature, while blue data points indicate lower values for the particular feature. A higher feature impact score indicates a higher contribution to a “severe” case prediction, while a lower feature impact score indicates a higher contribution to a “moderate” case prediction value.

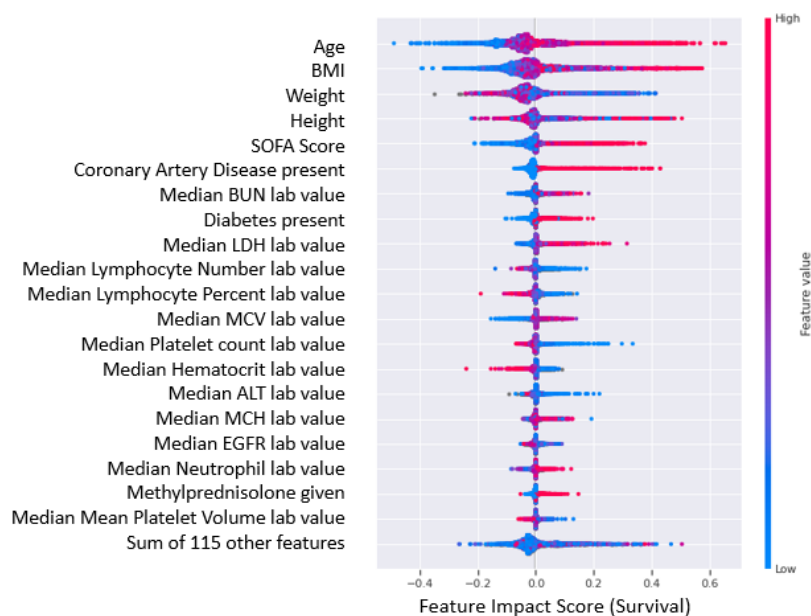


Figure 7. Top 20 most impactful features for COVID-19 survival predictive model. The red data points indicate higher values for the particular feature, while blue data points indicate lower values for the particular feature. A higher feature impact score indicates a higher contribution to a “no survival from COVID-19” prediction, while a lower feature impact score indicates a higher contribution to a “survival from COVID-19” prediction value.

Beyond comorbidities and medications, the following features, included biomarkers, have a specific pattern associated with severe COVID-19 cases and lower survival predicted by our AI/ML driven platform were further validated by external studies:

- Sequential Organ Failure Assessment (SOFA) Score: Higher SOFA score values are associated with more severe cases and lower survival [27].
- Lactate Dehydrogenase (LDH): Higher presence of LDH is associated with more severe cases and lower survival [28].
- Blood Urea Nitrogen (BUN): Higher presence of BUN is associated with more severe cases and lower survival especially with lower Serum Creatinine levels [29] and lower Albumin levels [30].

The following graphs shown in Figure 8 outline the specific range of values within scatter plots for impactful features as they pertain to COVID-19 severity.

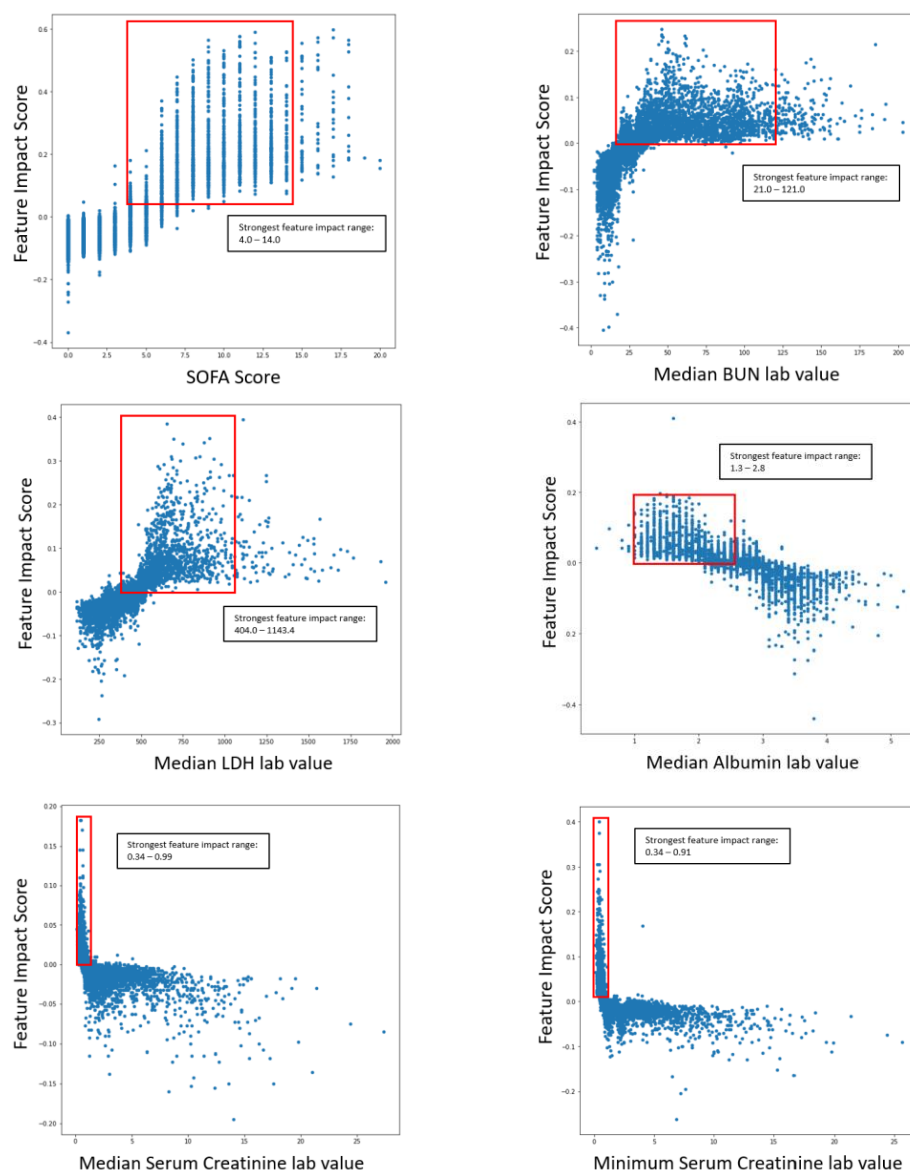


Figure 8. Range of impactful and validated biomarker values for COVID-19 severity prediction.

The ranges of values were calculated by measuring the 5th and 95th quantile of biomarker values when the severity was high. The same logic was applied to survival models. The following graphs in Figure 9 outline the specific range of values within scatter plots for impactful features concerning COVID-19 survival.

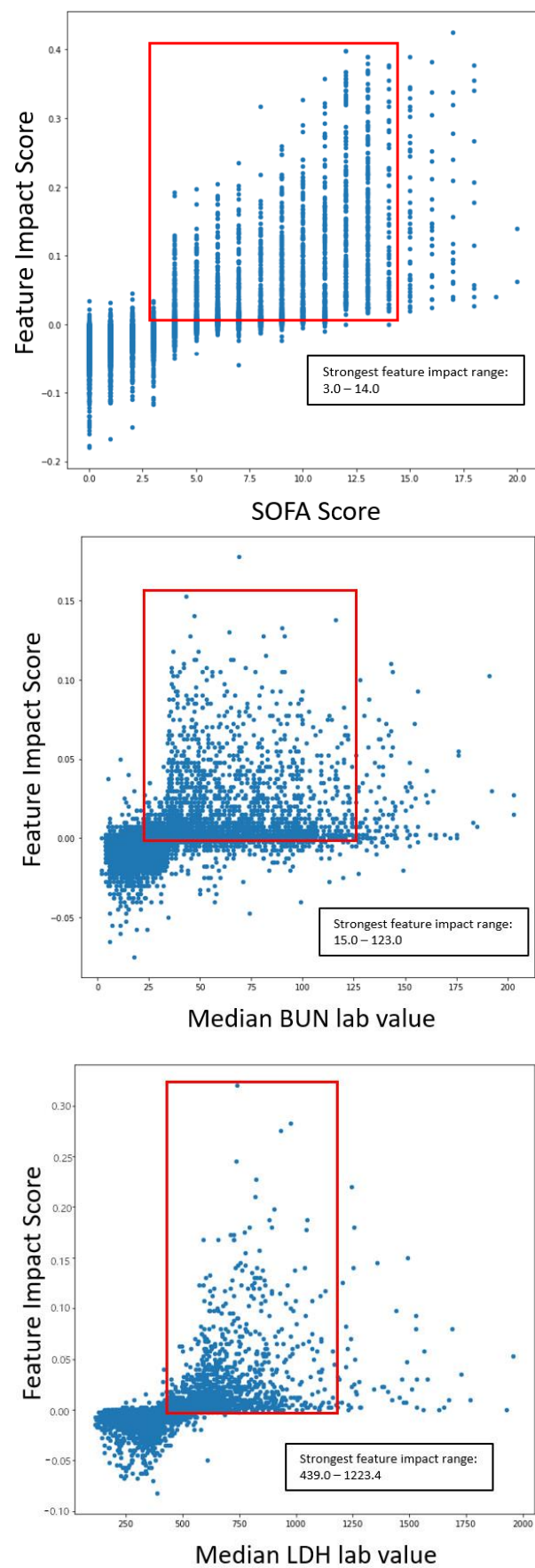


Figure 9. Range of impactful and validated biomarker values for COVID-19 survival prediction.

The scatter plots indicated the clear correlation between certain biomarker ranges and feature impact on each respective model. These validated features were selected for the gene correlation analysis to see what specific group of genes were associated with the upregulation of the given features.

With the generated SHAP values (feature impact scores) for each given feature value, a clustering analysis was performed to see how patients were grouped with varying biomarker levels. The clustering analysis was performed on feature impact scores based on the COVID-19 severity model explainability analysis. The elbow method [31] was used to determine the optimal number of clusters, and K-Means clustering was the clustering modality that generated the final clusters. Observations for each cluster and biomarker box plots are shown in (Supplementary Table S3, Supplementary Figure S1). It is worth noting that the distinction between the clusters were due to varying levels of the body weight, SOFA score, BUN, serum creatinine, LDH, albumin, PCO2 arterial, potassium levels, and lymphocyte count percent. Despite not showing any significant biomarker variation compared to the norm, Cluster 5 had a high percentage of severe COVID-19 cases among the patient group, with 93.1% affected. This suggests several less impactful biomarkers may have influenced the cluster's formation.

2.2.2. Omics Data Analysis of COVID19 Patients

Several gene co-expression modules were identified through gene network analysis, with MEcyan and MEDarkred standing out for their strong correlation with several key traits. Of these two modules, MEcyan was particularly noteworthy, as it showed a significant positive correlation with COVID-19 severity, end of organ damage (EOD), clinical biomarkers such as BUN, serum creatine, D-Dimer, creatine kinase, LDH, and SOFA score evaluation of patients. This finding aligns with the ML prediction referenced in Figure 10. However, only MEDarkred indicates a moderately significant association with comorbidities such as chronic kidney disease (CRD), heart failure, and alcohol/non-alcohol liver disease comorbidity in patients, as shown in Figure 11. In this study, MEcyan and MEDarkred gene modules were selected for functional annotation analysis and also analyzed for biological significance in the pathways across patients' traits, COVID-19 severity, patient disease comorbidity, and major clinical biomarker record of the patients.

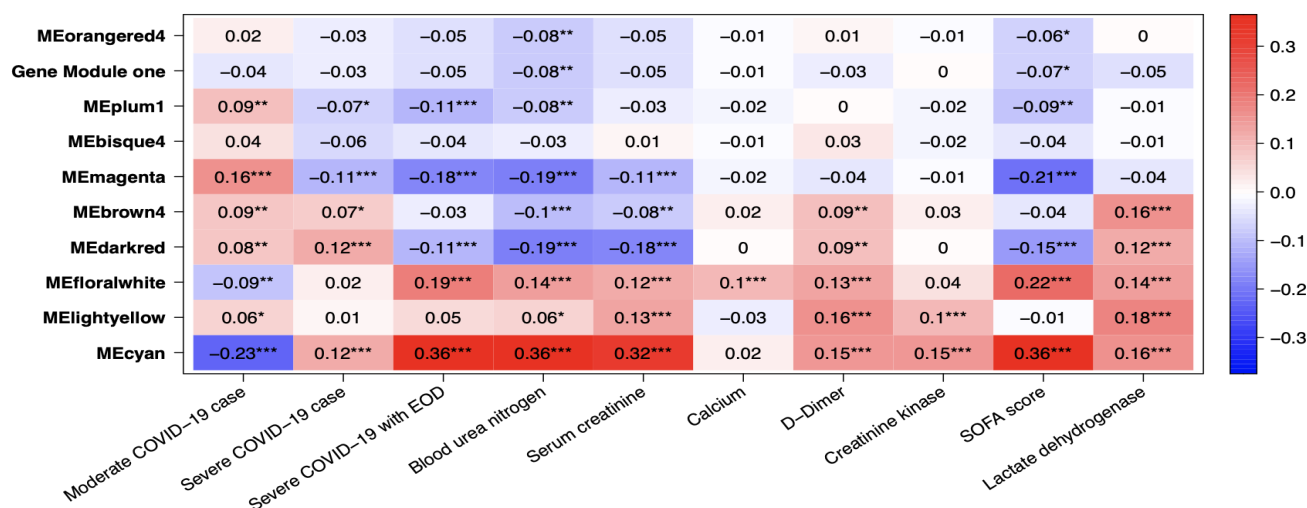


Figure 10. Heatmaps of module traits for severity and clinical biomarkers. The boxes indicate the correlation based on module eigengenes in the rows and traits in the column. The color legend—blue (negative correlation) and red (positive correlation). *p*-values represented by asterisks indicated significance. Asterisk signs indicate the order of significance where ***—very significant, **—significant, and *—significant.

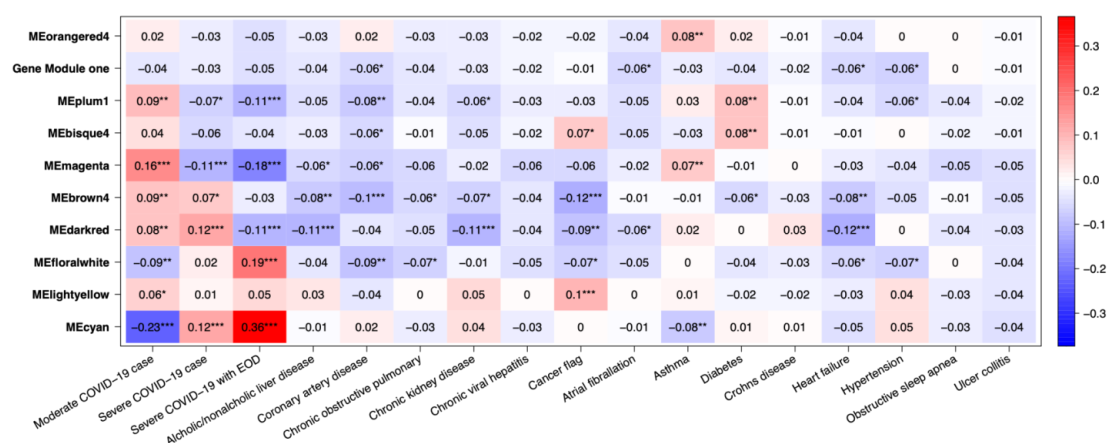


Figure 11. Heatmaps of module traits for severity and disease comorbidity. The boxes indicate the correlation based on module eigengenes in the rows and traits in the column. The color legend—blue (negative correlation) and red (positive correlation). *p*-values represented by asterisks indicated significance. Asterisk signs indicate the order of significance where ***—very significant, **—significant, and *—significant.

The genes present in the MEcyan module were subjected to functional enrichment analysis, which revealed their enrichment in pathways related to inflammatory and diverse immune responses, as per the Gene Ontology (GO) enrichment annotation (Figure 12). Additionally, they were found to be enriched in signaling pathways such as Tumor Necrosis Factor (TNF) signaling, cytokine signaling, Toll-like receptor signaling, and Interleukin-17 (IL-17) signaling pathway, based on the KEGG pathway (Figure 13). Additionally, several investigations have also confirmed our AI/ML-based predictive outcomes, demonstrating that these gene clusters were closely associated with immune signaling and response pathways that help combat viral infections [32–34]. The MEdarkred module's overrepresented pathways included platelet activation, regulation and response to blood coagulation, wound healing pathway, cell motility, and hemostasis for its genes. Those were considered important as the gene module results showed negative correlations with heart failure, chronic artery disease, and liver disease in patients (Figures 11 and 14).

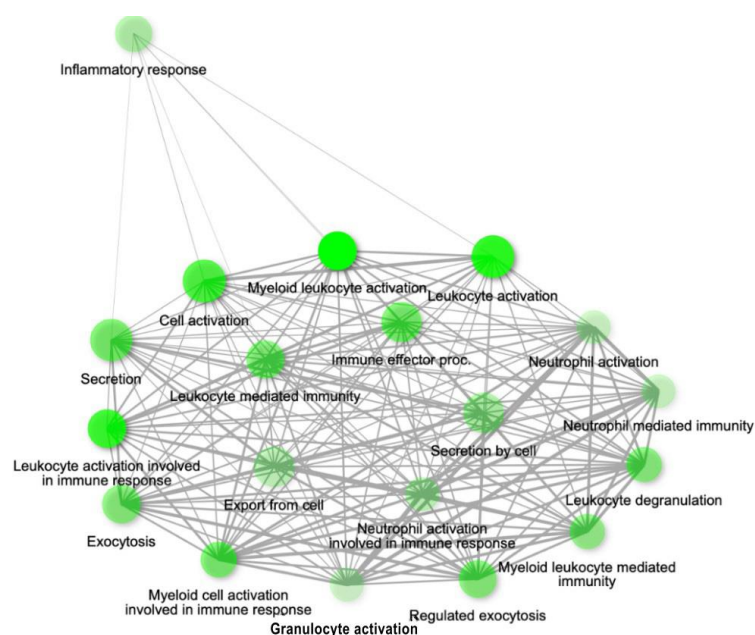


Figure 12. Gene network for top 20 enriched pathways for genes module, MEcyan based on GO biological process annotation.

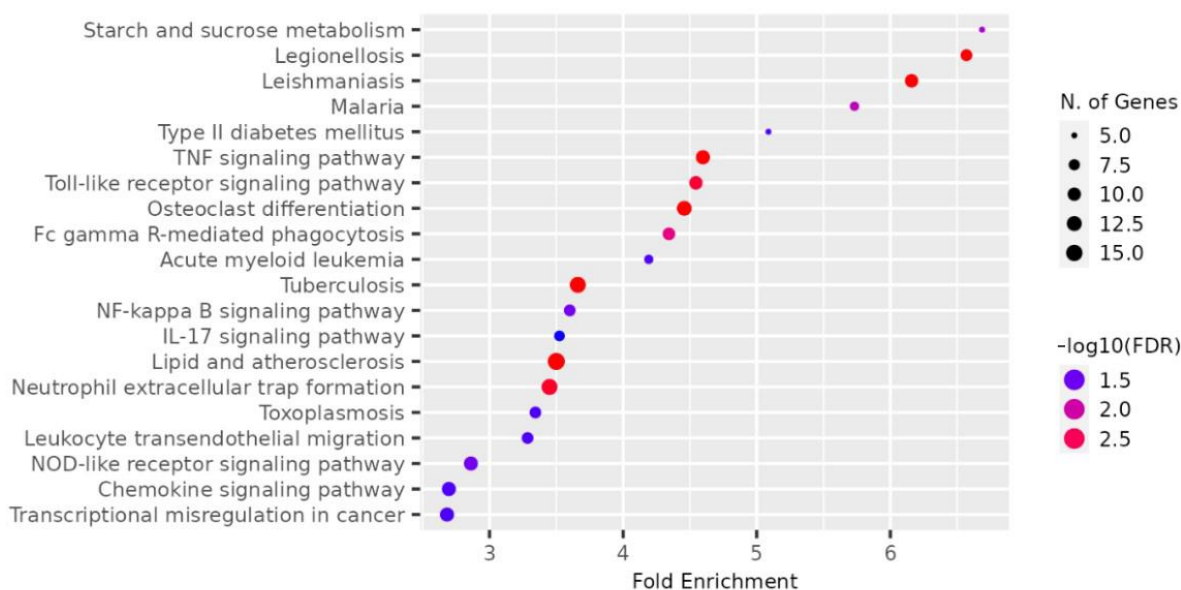


Figure 13. Dot plot for top 20 enriched pathways based for gene module, MEcyan based on KEGG.

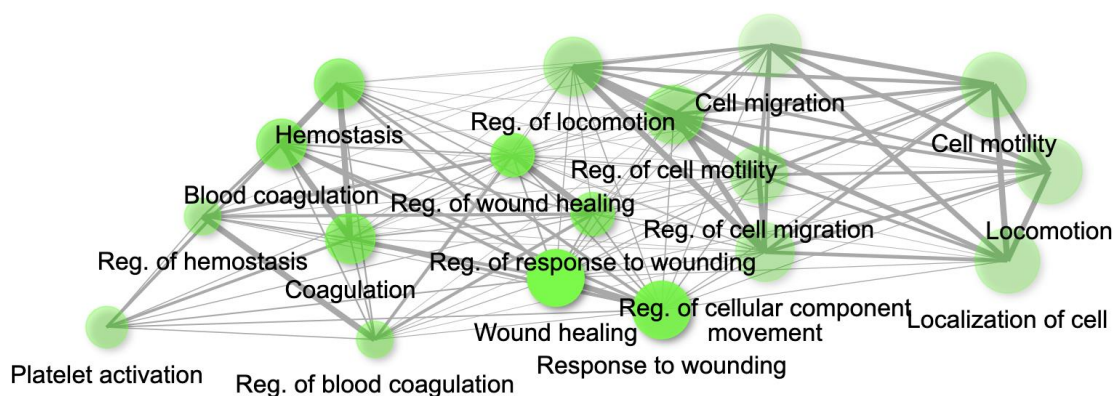


Figure 14. Gene network for top 20 enriched pathways for genes module, MEDarkred based on GO biological process annotation.

The estimated GS and MM genes are indications of how biologically significant such genes are in the enriched pathways in relation to the traits of interest (Supplementary Table S4). Our findings suggest that MEcyan modules harbor genes that exhibit a favorable association with COVID-19 severity markers, such as EOD, SOFA score, LDH, and BUN, while showing an unfavorable association with comorbidities such as heart failure and kidney disease (Supplementary Figure S7). MEDarkred modules show genes with a negative correlation with EOD, BUN, and SOFA, but a positive correlation with LDH. By applying our AI/ML modeling, based on both high significance and high intramodular connectivity, several noteworthy biomarkers were identified in the MEcyan and MEDarkred modules that are associated with traits such as EOD, SOFA, LDH, with GS > 0.3, and MM > 0.8. These gene-biomarkers (Supplementary Table S5) could be investigated in more detail for their potential use as drug targets or diagnostic tools. In addition, the analysis of patient OMICS datasets using AI/ML modeling has identified specific genes, including P2Y12, ECE1, MSANTD3-TMEFFI, PLEKHA8P1, NUTF2, SAV1, CXCR2P1, and MSANTD3, within the MEDarkred gene module that exhibit high GS and MM for clinical biomarkers such as SOFA score, BUN, EOD, and serum creatine. These genes are involved in biological pathways associated with blood coagulation and wound healing and have been implicated in the regulation of clinical biomarkers in patients with COVID-19.

3. Discussion

We have developed a robust AI/ML-based model that links OMICS datasets to clinical biomarkers for the stratification of patients and accurate predictions of disease severity and survival in a given cohort of COVID-19 patients. Given the high ROC-AUC score for both models, it can be inferred that the clinical features were sufficient to be an indicator for both COVID-19 severity and survival. However, the impact of each feature on the disease development and severity remains unclear. ML modeling was proposed as one of the promising approaches to identify the role of biomarker pathways in infection disease development. Certain restrictions on the dataset brought up potential model uncertainties as well. Sparsity in lab biomarker data restricts complete multivariate analysis for model explainability, as there would be a possibility for even more granular feature observations if lab biomarker data were more complete. There is a possibility that certain biomarkers, such as Hemoglobin A1C and Interleukin 8, which had more than 90% missing values, may have contributed to a significant impact on model predictability. The lack of diversity among observed comorbidities within patients in the dataset may have limited the potential observation for certain comorbidities to contribute to COVID-19 severity and survival based on model conclusions.

ML has been recognized as an insightful utility to predict the severity and survival of patients with COVID-19. A study published by Raman et al. in 2023 [35] highlighted the use of ML models to predict COVID-19 severity among 1795 patients from the University of Texas Southwest Medical Center. The study showed high predictive performance (ROC-AUC = 0.81, 0.82) among the trained models and highlighted what the trained models ranked as the most impactful biomarkers. There was no indication of biomarker data sparsity in the dataset used in the study. The models within the current analysis are much more dynamic as the complete set of biomarker features are not required per patient when predicting COVID-19 disease severity or survival. Additionally, the application of SHAP within the current analysis allowed for the exploration of another dimension within the features, which gave insights to specific feature value impact on model performance rather than indicating that given features may be important. In a study published by Ikemura et al., SHAP was employed in ML models that predicted COVID-19 mortality, which also achieved high model predictive performance (ROC-AUC = 0.803) [36]. However, the dataset within the study contained only 48 patient variables compared to 135 within the current analysis—leading to a larger variety of feature insights. The high model performance as well as considerations for missing data and the abundance of biomarker data makes the ML approach for current analysis highly dynamic and scalable, especially in a real-world scenario where patient data can be integrated to generate predictions.

SHAP was used to determine the most impactful clinical features and feature value ranges contributing highly to COVID-19 severity and survival. To determine the significant biomarkers for each use case, the top 20 most influential features were identified. It is noteworthy that age, body weight, and BMI were among the top 20 impactful features for both cases, indicating that clinicians should prioritize patients falling within the high impact ranges for these categories. Comorbidities for coronary artery disease and diabetes were seen within the top 20 impactful clinical features (conditions) for survival. However, comorbidities did not seem to have a significant impact when predicting COVID-19 severity, as they were not in the top 20 factors influencing disease severity. It is possible that comorbidities could exacerbate COVID-19, resulting in fatalities even in cases where the severity of the illness is not high. Surprisingly, the presence of methylprednisolone as a medication for patients with COVID-19 was seen as an impactful clinical feature for both severity and survival. There is no conclusive evidence to suggest that the medication itself increased severity or reduced survival. Subsequently, this may suggest the low efficiency of steroid therapy in COVID-19 patients. Therefore, it is probable that the medication's effect on certain biological pathways differs in different patient cohorts, which may influence significant biomarkers modulations found among this investigation. Another group of clinical features present in the top 20 impactful features for both models were blood

biomarkers, namely BUN, LDH, serum creatinine, and albumin. In addition, an SOFA score indicating a clinical organ damage measure was seen as a highly impactful feature for both use cases as well. The clinical biomarkers and SOFA score impact was aligned to the gene expression analysis, where highly correlated gene groups were identified in the dysregulation of clinical biomarkers in COVID-19 patients.

The most impactful SOFA score range for COVID-19 severity and mortality was 4–14 and 3–14, respectively. These ranges tie closely with observations seen in [27], which suggests that patients with SOFA scores greater than or equal to 5 have a higher risk of mortality. In regard to LDH [28] reports, patients with severe COVID-19 have greater than 350 U/L for their mean concentration, which is slightly less than the lower value of the most impactful median LDH concentration range of 404 and 439 for severity and mortality, respectively. A larger ratio between BUN and creatinine levels has shown to be an indicator for COVID-19 severity with optimal thresholds at 33.5 and 51.7 according to [29]. The most impactful range for both BUN and serum creatinine, according to SHAP analysis, suggests that the most impactful BUN/creatinine ratio is between 21 and 355, which is a larger range window than the study suggests. This wide range is likely due to the sparsity of biomarker data and the independent contributions of each BUN and serum creatinine to the model predictions; however, the higher impact on severity and survival seen for higher BUN and lower serum creatinine values does suggest that a larger ratio would be an indicator for both use cases. Similarly, a larger ratio between BUN and albumin levels has shown to be an indicator for COVID-19 severity with optimal thresholds greater than 3.9 according to [30], and SHAP analysis suggests that the most impactful range of BUN/albumin is between 7.5 and 93. The literature validates the usefulness of the BUN/creatinine and BUN/albumin ratios. However, to enhance the value further, it would be beneficial to include additional features in the data. This could involve calculating the actual BUN to serum creatinine and BUN to albumin ratios, along with incorporating more laboratory biomarker data.

We conducted a weight co-expression gene network analysis in addition to our ML prediction. This allowed us to approximate the patient gene expression and pinpoint the genes that display a noteworthy correlation with COVID-19 severity, disease comorbidity, and the most critical clinical biomarkers. Our predictive analysis results suggest that there is functional enrichment in pathways related to inflammation, cytokine response, and Toll-like and interleukin signaling in the genes network that are correlated positively with COVID-19 patients having high BUN, LDH, serum creatinine levels, an elevated SOFA score, and EOD. Within the MEcyan modules, several crucial genes display characteristics such as interleukin receptors (IL8RBP, IL10RB, IL17RA) and Toll-like receptors (TLR1, TLR5, TLR4, TLR6, TLR8, TLR2) alongside MYD88. These genes are intricately associated with activating pro-inflammatory cytokines, signaling pathways, and sensing the SARS-COV-2 envelope protein [34,37]. In addition, TLR2 and MYD88 have been shown to be associated with COVID-19 disease severity. It is important to acknowledge that cytokines, despite being integral to the innate immune response to viral infections, can cause significant harm to organs and tissues when dysregulated or overly inflamed, leading to the onset of cytokine storms [3–34]. Hence, patients with hyperinflammation are often placed on immunosuppression, immunomodulation, and selective cytokine blocking drugs to improve their outcome [32,38,39]. Additionally, interleukin 10 receptor, IL10RB, was identified to be significant to patients' severity and has been previously noted as a regulator of host susceptibility to COVID-19 severity and potential drug target [40,41].

Furthermore, the set of significant genes represented in the MEdarkred module (Supplementary Table S10) across EOD, SOFA, BUN, and heart failure, alcohol liver disease, and kidney disease comorbidity include gene markers such as AMAD9, KIF1B, SLC22A4, MAPK14, MAP2K6, SLC2A3, ILI7RA, which have been associated with COVID-19 disease severity, tissue injury, and sepsis. For instance, an increased expression of ADAM9 and others member of the metalloprotease protein family in the blood has been associated with tissue damage and identified as a potential drug target for COVID-19 [42–47]. Rinchai et al.

particularly highlight the involvement of AMAD9 metalloprotease in cellular adhesion, protein shedding, various cell population and activity including cell–cell interaction, and fibroblast activity—a key part of connective tissue and extracellular matrix (ECM) formation. This is consistent with our findings where the annotation of the MEDarkred gene module is shown to be enriched in cellular component (Supplementary Figures S8 and S9) activities such as the cell surface, focal adhesion, cell cortex, anchoring junction, stress fiber, platelet activities, myofibrils, and actin cytoskeleton. These networks of biological and cellular activities further emphasize the importance of better understanding observed abnormalities in the extracellular matrix (ECM), inflammatory response, activities of metalloproteinase proteins, and other key significant gene biomarkers in COVID-19 patients.

Our AI/ML platform demonstrated that most disease comorbidities exhibited minimal correlation, except for alcohol/non-alcoholic liver disease, diabetes, chronic kidney disease, and heart failure, which showed moderately significant negative correlation in the co-expression MEDarkred gene modules. These genes were enriched in pathways related to wound healing, blood coagulation, hemostasis, and cell motility, as shown in Figures 11 and 14. This biological insight strengthens our previous observations regarding the link between hyper-inflammation, cytokine reactions, cytokine storms, and tissue and organ damage. This understanding may clarify the moderate correlation observed in gene modules for chronic kidney disease, heart failure, and liver disease with functional annotation enriched in wound healing, hemostasis, blood clotting control, and coagulation. Identifying significant gene-clinical biomarkers with high GS and MM scores can facilitate the selection of key genes for potential drug targeting based on their repositioning and druggability in the context of personalized medicine.

4. Materials and Methods

4.1. Clinical Data Acquisition

An analytical review of the literature indexed in the PubMed/MEDLINE and Scopus databases was conducted. The goal was to summarize all available sources providing clinical and OMICs data for individual patients infected with SARS-CoV-2 and admitted to the healthcare institutions. To identify clinical publications relevant to COVID-19 patients with appropriate datasets, a search strategy was employed using both broad terms and specific biomarker terminology to maximize the number of identified publications. We included peer-reviewed, pre-proof, and papers published ahead of print that reported COVID-19 cases, confirmed using a real-time reverse transcriptase-polymerase chain reaction. Articles in English and non-English languages were selected. The first run of the literature search using general terminology returned more than three hundred thousand article titles. Following, the search ranking algorithm was applied to triage the sources found, and twenty-two research articles with the primary patient data (Supplementary Data S1) were evaluated and used as a database for further ML training.

Specific sets of clinical terms with assigned logical weight for biomarker identification were deployed in the AI algorithm to further rank the source according to their relevance in providing individual clinical data.

4.2. Data Curation

The clinical dataset consisted of patient conditions, lab test results, and clinician reports including a Sequential Organ Failure Assessment (SOFA) score, which is used to predict ICU mortality based on lab results and clinical data. A SOFA score is considered helpful for ML model training because it directly reflects the patient's condition based on the oxygen saturation, respiratory function, platelets count, blood pressure values, creatinine, and bilirubin test values as well as the Glasgow coma scale. SOFA score has demonstrated its reliability during its use for more than 25 years since it was developed [48]. The full clinical datasets for these patients were downloaded from Synapse with the identification number syn35874390 from synapse.org. The dataset was filtered to include patients that had tested positive for COVID-19 through either a PCR and/or antibody tests. There were 581 unique patients with a combined set of 7707 clinical data points containing multiple days of patient

information. Individual patient parameters were categorized into demographic parameters, comorbidities, blood cell count parameters, and biochemical and inflammatory biomarkers. Duplicates were dropped to result in the overall 7707 clinical data points to ensure that unique patient information would be used to train the machine learning model. The gene expression dataset for this analysis was obtained from the Gene Expression Omnibus (GEO) with the accession number GSE215865. The dataset consisted of the transcriptomics (RNA-seq) of whole blood samples of hospitalized COVID-19 patients and healthy, hospitalized control patients [49]. Using the metadata available, the gene expression was reduced to 1198 samples after removing sets of patients with largely incomplete clinical descriptions [50,51].

4.3. Bioinformatics Methodology

Gene network analysis was conducted on 1198 gene expression data with the WGCNA package [52] to identify significant correlation between co-expressed gene modules and patients' disease severity, comorbidity, and clinical biomarkers. First, differentially expressed genes (DEGs) between COVID-19-positive and COVID-19-negative (control) patients were extracted using normalized gene counts (at p -value < 0.05 and $\text{abs}(\log\text{FC}) \geq 1$) using limma package [53], and fed into construction of weighted gene co-expression networks. Using WGCNA functions, approximate scale free topology was determined to define the adjacency matrix and transformed into topological overlaps (TOM) and dissimilarity (disTOM) matrix. A hierarchical clustering was performed to identify gene modules with the minimum size set at 30 genes. Modules with gene expression similarity were subsequently merged based on module eigengene and cluster correlation. Additionally, to determine key genes associated with COVID-19 severity and other clinical features, Gene Significance (GS) and Module Membership was calculated for every gene in the modules and filtered at $\text{GS} > 0.2$ and $\text{MM} > 0.8$. Genes with high values typically have high connection and interaction, and do have strong associations with traits of interests [52]. The selected key gene biomarkers are used for further analysis for biological annotations and activities. To gain insights into the biological role of these genes, functional annotation analysis was performed using Gene Ontology (GO) [54,55] and Kyoto Encyclopedia of Genes and Genome (KEGG) [56] using ShuyGO web-based tools [57].

4.4. Descriptor Analysis and Selection

For the initial dataset used in model training, we collected various types of data such as patient condition, biomarkers, comorbidities, and therapy information. The data were classified as either numerical or categorical data. The dataset also had biomarkers with varying levels of missing values, each of which was identified in the missing values table (Supplementary Table S2). Each relevant column used for model training is indicated in the column groups table (Supplementary Table S1).

All features were then normalized and evaluated to discern relative importance to the respective target features (survival outcome and disease severity) using our ML infrastructure pipeline. These algorithms were optimized to the data available in the training set. Algorithm performance in selecting essential features were proportional to the training data size [58]. BIOiSIM™'s AI infrastructure included automated statistical learning algorithms prioritizing a large collection of features and selecting those features that improved model performance. This subset was further refined during model training through Recursive Feature Elimination (RFE) [59] and Boruta selection algorithms [60].

4.5. Model Training and Evaluation

We developed two different types of ML models for this project. Both classification models used the biomarkers information, comorbidities, and therapy information to predict either COVID-19 case severity or survival outcome.

Inbuilt MLs were used for both classification and regression models. The ML algorithms used in the study included linear algorithms such as the Bayesian ridge, support vector machines, tree-based ensemble methods such as LightBoost, CatBoost, and XG-

Boost, random forests, and multi-layer perceptions [61]. Features selected from the initially assembled descriptor set were used as inputs in the training process.

All individual patient data contained information regarding the survival outcomes and length of hospital stay as the initial set for model training and evaluation. They were considered critical for model learning. The ML infrastructure randomly divided the entire dataset into 80% and 20% portions for the training and test splits. Hyperparameter optimization was conducted on the validation sets obtained through 10-fold stratified cross-validation of the training set, which involved creating validation sets using stratified splits. Model evaluation metrics, available in Section 2, were determined for the final unseen test holdout set, which was distinct from the training and cross-fold validation sets used in model training.

Classification model evaluation metrics include balanced accuracy and a ROC-AUC score (Area Under the Receiver Operating Characteristic Curve). Balanced accuracy is defined as an optimal approach to the standard accuracy metric, which is adjusted to perform better on imbalanced datasets. The ROC-AUC score assesses the distinction of probability predictions between two classes such as severe COVID-19 or moderate in the context of severity predictions. Balanced accuracy was selected since the survival model training dataset was imbalanced and to maintain consistent evaluation with the severity models. The ROC-AUC evaluation enables for the assessment of the predicted probabilities, which could be advantageous if the models were to be utilized in a clinical environment. This approach provides probabilistic information regarding the status of a patient with COVID-19. Subsequently, both metrics were determined on each test set from each respective trained model.

5. Conclusions

A robust AI/ML-based model was created to stratify COVID-19 patients using OMICS, and clinical biomarker datasets, which enables accurate prediction of disease severity and outcomes. The accuracy of both models was 98.1% and 99.9%, respectively. The stratification module is based on the ML/DL algorithms with built-in classification and regression models. A unified dataset comprising of clinical biomarkers and the OMICS dataset were used as primary inputs, and their multi-dimensional unified analysis helped identify, with high accuracy, critical transcriptomic and clinical biomarkers for the prediction of disease severity and outcomes. Our AI/ML-driven platform also demonstrated that comorbidities such as alcohol/non-alcoholic liver disorders, diabetes, chronic kidney disease, and heart failure could impact the severity of the disease and the outcomes of COVID-19 patients. The genomic analysis of the patients as a whole revealed the pathways implicated in the development of the disease along with related clinical biomarkers. Our research has demonstrated that patient stratification models, driven by AI/ML modeling, can be expanded to other viral infections, and could be used to precisely identify the manifestation of clinical biomarkers, resulting in more accurate diagnoses and treatment options in the context of personalized medicine.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ijms24076250/s1>.

Author Contributions: B.B.—data curation and evaluation, data analysis; Y.N.B.—development of the ML model and generation of the simulation results; R.B.—development of the ML models, data visualization, and ML explainability analysis; M.K.—study design, evaluation of the results, text writing and submission of the manuscript; S.W.B.—scientific accuracy and advising; K.C.—conceptualization of the study and clinical data review; J.V.—general project management. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in the study were received from publicly available sources; the list of sources with the references is provided in the Supplementary Data S1.

Conflicts of Interest: The authors are all employees of VeriSIM Life and used a proprietary AI/ML-driven platform to generate the outcomes for the manuscript.

References

- Lee, B.; Lewis, G.; Agyei-Manu, E.; Atkins, N.; Bhattacharyya, U.; Dozier, M.; Rostron, J.; Sheikh, A.; McQuillan, R.; Theodoratou, E.; et al. Risk of serious COVID-19 outcomes among adults and children with moderate-to-severe asthma: A systematic review and meta-analysis. *Eur. Respir. Rev.* **2022**, *31*, 220066. [\[CrossRef\]](#) [\[PubMed\]](#)
- Song, C.Y.; Xu, J.; He, J.Q.; Lu, Y.Q. Immune dysfunction following COVID-19, especially in severe patients. *Sci. Rep.* **2020**, *10*, 15838. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, J.; Yang, W.; Huang, F.; Zhang, K. Effectiveness of mRNA and viral-vector vaccines in epidemic period led by different SARS-CoV-2 variants: A systematic review and meta-analysis. *J. Med. Virol.* **2023**, *95*, e28623. [\[CrossRef\]](#) [\[PubMed\]](#)
- Whiteside, T.; Kane, E.; Aljohani, B.; Alsamman, M.; Pourmand, A. Redesigning emergency department operations amidst a viral pandemic. *Am. J. Emerg. Med.* **2020**, *38*, 1448–1453. [\[CrossRef\]](#)
- Sipior, J.C. Considerations for development and use of AI in response to COVID-19. *Int. J. Inf. Manag.* **2020**, *55*, 102170. [\[CrossRef\]](#)
- Vo, T.H.; Nguyen, N.T.K.; Kha, Q.H.; Le, N.Q.K. On the road to explainable AI in drug-drug interactions prediction: A systematic review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2112–2123. [\[CrossRef\]](#)
- Basu, K.; Sinha, R.; Ong, A.; Basu, T. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? *Indian J. Dermatol.* **2020**, *65*, 365–370. [\[CrossRef\]](#)
- Lam, C.; Calvert, J.; Siefkas, A.; Barnes, G.; Pellegrini, E.; Green-Saxena, A.; Hoffman, J.; Mao, Q.; Das, R. Personalized stratification of hospitalization risk amidst COVID-19: A machine learning approach. *Health Policy Technol.* **2021**, *10*, 100554. [\[CrossRef\]](#)
- Hartzband, P.; Groopman, J. Physician Burnout, Interrupted. *N. Engl. J. Med.* **2020**, *382*, 2485–2487. [\[CrossRef\]](#)
- Kannampallil, T.G.; Goss, C.W.; Evanoff, B.A.; Strickland, J.R.; McAlister, R.P.; Duncan, J. Exposure to COVID-19 patients increases physician trainee stress and burnout. *PLoS ONE* **2020**, *15*, e0237301. [\[CrossRef\]](#)
- Zhang, K.; Liu, X.; Shen, J.; Li, Z.; Sang, Y.; Wu, X.; Zha, Y.; Liang, W.; Wang, C.; Wang, K.; et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell* **2020**, *182*, 1360. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liang, W.; Liang, H.; Ou, L.; Chen, B.; Chen, A.; Li, C.; Li, Y.; Guan, W.; Sang, L.; Lu, J.; et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients with COVID-19. *JAMA Intern. Med.* **2020**, *180*, 1081–1089. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, Z.; Su, C.; Xiao, Y.; Wang, F. Artificial intelligence for COVID-19: Battling the pandemic with computational intelligence. *Intell. Med.* **2022**, *2*, 13–29. [\[CrossRef\]](#) [\[PubMed\]](#)
- Williams, R.D.; Markus, A.F.; Yang, C.; Salles, T.D.; Falconer, T.; Jonnagaddala, J.; Kim, C.; Rho, Y.; Williams, A.; An, M.H.; et al. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv* **2020**, *27*, 2020-05. [\[CrossRef\]](#)
- Rapsang, A.G.; Shyam, D.C. Scoring systems in the intensive care unit: A compendium. *Indian J. Crit. Care Med.* **2014**, *18*, 220–228. [\[CrossRef\]](#)
- Fan, G.; Tu, C.; Zhou, F.; Liu, Z.; Wang, Y.; Song, B.; Gu, X.; Wang, Y.; Wei, Y.; Li, H.; et al. Comparison of severity scores for COVID-19 patients with pneumonia: A retrospective study. *Eur. Respir. J.* **2020**, *56*, 2002113. [\[CrossRef\]](#)
- Su, Y.; Tu, G.W.; Ju, M.J.; Yu, S.J.; Zheng, J.L.; Ma, G.G.; Liu, K.; Ma, J.F.; Yu, K.H.; Xue, Y.; et al. Comparison of CRB-65 and quick sepsis-related organ failure assessment for predicting the need for intensive respiratory or vasopressor support in patients with COVID-19. *J. Infect.* **2020**, *81*, 647–679. [\[CrossRef\]](#)
- Yang, H.S.; Hou, Y.; Vasovic, L.V.; Steel, P.A.D.; Chadburn, A.; Racine-Brzostek, S.E.; Velu, P.; Cushing, M.M.; Loda, M.; Kaushal, R.; et al. Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning. *Clin. Chem.* **2020**, *66*, 1396–1404. [\[CrossRef\]](#)
- Calzari, L.; Zanolli, L.; Inglese, E.; Scaglione, F.; Cavagnola, R.; Ranucci, F.; Di Blasio, A.M.; Stefanini, G.; Carlo, G.; Parati, G.; et al. Role of epigenetics in the clinical evolution of COVID-19 disease. Epigenome-wide association study identifies markers of severe outcome. *Eur. J. Med. Res.* **2023**, *28*, 81. [\[CrossRef\]](#)
- Tsiftoglou, S.A.; Gavrilaki, E.; Touloumenidou, T.; Koravou, E.E.; Koutra, M.; Papayanni, P.G.; Karali, V.; Papalexandri, A.; Varelas, C.; Chatzopoulou, F.; et al. Targeted genotyping of COVID-19 patients reveals a signature of complement C3 and factor B coding SNPs associated with severe infection. *Immunobiology* **2023**, *228*, 152351. [\[CrossRef\]](#)
- Moreno, R.; Rhodes, A.; Piquilloud, L.; Hernandez, G.; Takala, J.; Gershengorn, H.B.; Tavares, M.; Coopersmith, C.M.; Myatra, S.N.; Singer, M.; et al. The Sequential Organ Failure Assessment (SOFA) Score: Has the time come for an update? *Crit. Care* **2023**, *27*, 15. [\[CrossRef\]](#) [\[PubMed\]](#)
- Azur, M.J.; Stuart, E.A.; Frangakis, C.; Leaf, P.J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [\[CrossRef\]](#) [\[PubMed\]](#)

23. Khosravi, P.; Vergari, A.; Choi, Y.; Liang, Y.; Broeck, G.V.D. Handling Missing Data in Decision Trees: A Probabilistic Approach. *arXiv* **2020**, arXiv:2006.16341. [[CrossRef](#)]
24. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
25. Guan, W.J.; Liang, W.H.; He, J.X.; Zhong, N.S. Cardiovascular comorbidity and its impact on patients with COVID-19. *Eur. Respir. J.* **2020**, *55*, 2001227. [[CrossRef](#)]
26. Ssentongo, P.; Ssentongo, A.E.; Heilbrunn, E.S.; Ba, D.M.; Chinchilli, V.M. Association of cardiovascular disease and 10 other pre-existing comorbidities with COVID-19 mortality: A systematic review and meta-analysis. *PLoS ONE* **2020**, *15*, e0238215. [[CrossRef](#)] [[PubMed](#)]
27. Chinta, S.; Rodriguez-Guerra, M.; Shaban, M.; Pandey, N.; Jaquez-Duran, M.; Vittorio, T.J. COVID-19 therapy and vaccination: A clinical narrative review. *Drugs Context* **2023**, *12*, 1–11. [[CrossRef](#)]
28. Yang, Z.; Hu, Q.; Huang, F.; Xiong, S.; Sun, Y. The prognostic value of the SOFA score in patients with COVID-19: A retrospective, observational study. *Medicine* **2021**, *100*, e26900. [[CrossRef](#)]
29. Li, C.; Ye, J.; Chen, Q.; Hu, W.; Wang, L.; Fan, Y.; Lu, Z.; Chen, J.; Chen, Z.; Chen, S.; et al. Elevated Lactate Dehydrogenase (LDH) level as an independent risk factor for the severity and mortality of COVID-19. *Aging* **2020**, *12*, 15670–15681. [[CrossRef](#)]
30. Ok, F.; Erdogan, O.; Durmus, E.; Carkci, S.; Canik, A. Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients. *J. Med. Virol.* **2021**, *93*, 786–793. [[CrossRef](#)]
31. Kucukceran, K.; Ayranci, M.K.; Girisgin, A.S.; Kocak, S.; Dundar, Z.D. The role of the BUN/albumin ratio in predicting mortality in COVID-19 patients in the emergency department. *Am. J. Emerg. Med.* **2021**, *48*, 33–37. [[CrossRef](#)]
32. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [[CrossRef](#)]
33. Channappanavar, R.; Perlman, S. Pathogenic human coronavirus infections: Causes and consequences of cytokine storm and immunopathology. *Semin. Immunopathol.* **2017**, *39*, 529–539. [[CrossRef](#)] [[PubMed](#)]
34. 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](#)] [[PubMed](#)]
35. Zheng, M.; Karki, R.; Williams, E.P.; Yang, D.; Fitzpatrick, E.; Vogel, P.; Jonsson, C.B.; Kanneganti, T.D. TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. *Nat. Immunol.* **2021**, *22*, 829–838. [[CrossRef](#)]
36. Raman, G.; Ashraf, B.; Demir, Y.K.; Kershaw, C.D.; Cheruku, S.; Atis, M.; Atis, A.; Atar, M.; Chen, W.; Ibrahim, I.; et al. Machine learning prediction for COVID-19 disease severity at hospital admission. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 46. [[CrossRef](#)]
37. Ikemura, K.; Bellin, E.; Yagi, Y.; Billett, H.; Saada, M.; Simone, K.; Stahl, L.; Szymanski, J.; Goldstein, D.Y.; Reyes Gil, M. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. *J. Med. Internet Res.* **2021**, *23*, e23458. [[CrossRef](#)]
38. Pacha, O.; Sallman, M.A.; Evans, S.E. COVID-19: A case for inhibiting IL-17? *Nat. Rev. Immunol.* **2020**, *20*, 345–346. [[CrossRef](#)]
39. Shakoory, B.; Carcillo, J.A.; Chatham, W.W.; Amdur, R.L.; Zhao, H.; Dinarello, C.A.; Cron, R.Q.; Opal, S.M. Interleukin-1 Receptor Blockade Is Associated With Reduced Mortality in Sepsis Patients With Features of Macrophage Activation Syndrome: Reanalysis of a Prior Phase III Trial. *Crit. Care Med.* **2016**, *44*, 275–281. [[CrossRef](#)] [[PubMed](#)]
40. Wang, D.; Fu, B.; Peng, Z.; Yang, D.; Han, M.; Li, M.; Yang, Y.; Yang, T.; Sun, L.; Li, W.; et al. Tocilizumab in patients with moderate or severe COVID-19: A randomized, controlled, open-label, multicenter trial. *Front. Med.* **2021**, *15*, 486–494. [[CrossRef](#)]
41. Voloudakis, G.; Hoffman, G.; Venkatesh, S.; Lee, K.M.; Dobrindt, K.; Vicari, J.M.; Zhang, W.; Beckmann, N.D.; Jiang, S.; Hoagland, D.; et al. IL10RB as a key regulator of COVID-19 host susceptibility and severity. *medRxiv* **2021**. [[CrossRef](#)]
42. Voloudakis, G.; Vicari, J.M.; Venkatesh, S.; Hoffman, G.E.; Dobrindt, K.; Zhang, W.; Beckmann, N.D.; Higgins, C.A.; Argyriou, S.; Jiang, S.; et al. A translational genomics approach identifies IL10RB as the top candidate gene target for COVID-19 susceptibility. *NPJ Genom. Med.* **2022**, *7*, 52. [[CrossRef](#)] [[PubMed](#)]
43. 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](#)] [[PubMed](#)]
44. Cui, S.; Niu, K.; Xie, Y.; Li, S.; Zhu, W.; Yu, L.; Tan, H. Screening of potential key ferroptosis-related genes in sepsis. *PeerJ* **2022**, *10*, e13983. [[CrossRef](#)] [[PubMed](#)]
45. Fan, Y.; Han, Q.; Li, J.; Ye, G.; Zhang, X.; Xu, T.; Li, H. Revealing potential diagnostic gene biomarkers of septic shock based on machine learning analysis. *BMC Infect. Dis.* **2022**, *22*, 65. [[CrossRef](#)]
46. Huang, J.J.; Wang, C.W.; Liu, Y.; Zhang, Y.Y.; Yang, N.B.; Yu, Y.C.; Jiang, Q.; Song, Q.F.; Qian, G.Q. Role of the extracellular matrix in COVID-19. *World J. Clin. Cases* **2023**, *11*, 73–83. [[CrossRef](#)]
47. Rinchai, D.; Kewcharoenwong, C.; Kessler, B.; Lertmemongkolkhai, G.; Chaussabel, D. Increased abundance of ADAM9 transcripts in the blood is associated with tissue damage. *F1000Res* **2015**, *4*, 89. [[CrossRef](#)]
48. Saifi, M.A.; Bansod, S.; Godugu, C. COVID-19 and fibrosis: Mechanisms, clinical relevance, and future perspectives. *Drug Discov. Today* **2022**, *27*, 103345. [[CrossRef](#)]
49. Thompson, R.C.; Simons, N.W.; Wilkins, L.; Cheng, E.; Del Valle, D.M.; Hoffman, G.E.; Cervia, C.; Fennessy, B.; Mouskas, K.; Francoeur, N.J.; et al. Molecular states during acute COVID-19 reveal distinct etiologies of long-term sequelae. *Nat. Med.* **2023**, *29*, 236–246. [[CrossRef](#)]

50. Charney, A.W.; Simons, N.W.; Mouskas, K.; Lepow, L.; Cheng, E.; Le Berichel, J.; Chang, C.; Marvin, R.; Del Valle, D.M.; Calorossi, S.; et al. Sampling the host response to SARS-CoV-2 in hospitals under siege. *Nat. Med.* **2020**, *26*, 1157–1158. [[CrossRef](#)]
51. Laposata, M. (Ed.) Clinical Laboratory Reference Values. In *Laboratory Medicine: The Diagnosis of Disease in the Clinical Laboratory*; McGraw-Hill Education: New York, NY, USA, 2014.
52. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [[CrossRef](#)]
53. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids. Res.* **2015**, *43*, e47. [[CrossRef](#)] [[PubMed](#)]
54. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [[CrossRef](#)] [[PubMed](#)]
55. Gene Ontology Consortium. The Gene Ontology resource: Enriching a Gold mine. *Nucleic Acids. Res.* **2021**, *49*, D325–D334. [[CrossRef](#)]
56. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids. Res.* **2021**, *49*, D545–D551. [[CrossRef](#)]
57. Ge, S.X.; Jung, D.; Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, *36*, 2628–2629. [[CrossRef](#)]
58. Antontsev, V.; Jagarapu, A.; Bunday, Y.; Hou, H.; Khotimchenko, M.; Walsh, J.; Varshney, J. A hybrid modeling approach for assessing mechanistic models of small molecule partitioning in vivo using a machine learning-integrated modeling platform. *Sci. Rep.* **2021**, *11*, 11143. [[CrossRef](#)]
59. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [[CrossRef](#)]
60. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A System for Feature Selection. *Fundam. Inform.* **2010**, *101*, 271–285. [[CrossRef](#)]
61. Spooner, A.; Chen, E.; Sowmya, A.; Sachdev, P.; Kochan, N.A.; Trollor, J.; Brodaty, H. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Rep.* **2020**, *10*, 20410. [[CrossRef](#)]

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.