Evaluating biomedical feature fusion on machine learning's predictability and interpretability of COVID-19 severity types

Authors: Haleigh West-Page[1], Kevin McGoff[1], Harrison Latimer[1], Isaac Olufadewa[2], Shi Chen[2]

Affiliations: [1]University of North Carolina at Charlotte, Department of Mathematics and Statistics; [2]University of North Carolina at Charlotte, Department of Public Health Sciences

Abstract

Background: Accurately differentiating severe from non-severe COVID-19 clinical types is critical for the healthcare system to optimize workflow, as severe patients require intensive care. Current techniques lack the ability to accurately predict COVID-19 patients' clinical type, especially as SARS-CoV-2 continues to mutate.

Objective: In this work, we explore both predictability and interpretability of multiple state-of-the-art machine learning (ML) techniques trained and tested under different biomedical data types and COVID-19 variants.

Methods: Comprehensive patient-level data were collected from 362 patients (214 severe, 148 non-severe) with the original SARS-CoV-2 variant in 2020 and 1000 patients (500 severe, 500 non-severe) with the Omicron variant in 2022-2023. The data included 26 biochemical features from blood testing and 26 clinical features from each patient's clinical characteristics and medical history. Different types of ML techniques, including penalized logistic regression (LR), random forest (RF), *k*-nearest neighbors (kNN), and support vector machines (SVM) were applied to build predictive models based on each data modality separately and together for each variant set.

Results: All ML models performed similarly under different testing scenarios. The fused characteristic modality yielded the highest area under the curve (AUC) score achieving 0.914 on average. The second highest AUC was 0.876 achieved by the biochemical modality alone, followed by 0.825 achieved by clinical modality alone. All ML models were robust when cross-tested with original and Omicron variant patient data. Upon model interpretation, our models ranked elevated d-dimer (biochemical feature), elevated high sensitivity troponin I (biochemical feature), and age greater than 55 years (clinical feature) as the most predictive features of severe COVID-19.

Conclusions: We found ML to be a powerful tool for predicting severe COVID-19 based on comprehensive individual patient-level data. Further, ML models trained on the biochemical and clinical modalities together witness enhanced predictive power. The improved performance of these ML models when trained and cross-tested with Omicron variant data supports the robustness of ML as a tool for clinical decision support.

Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus has been impacting healthcare systems everywhere. As of February 4th, 2024, cumulative cases have exceeded 774 million, with more than 7 million deaths worldwide according to the World Health Organization (WHO) [18]. During these four years of the pandemic, several major SARS-CoV-2 variants and subvariants have manifested, with the Omicron variant being the most persistent since November 2021 [24]. Machine learning (ML) have made substantial contributions to various aspects of the pandemic, including diagnostic decision support [15], developing novel pharmaceutics [16], and predicting trajectories of the pandemic [17], among many others.

A less addressed yet critical effect of the pandemic has been the sudden increased burden on healthcare facilities, namely hospitals. The influx of severe COVID-19 patients overwhelms intensive care units (ICU), which results in increased mortality [1], especially in regions in the U.S. with less health resources (e.g., ICU and general hospital beds) across the pandemic ([14], [18]). These challenges require solutions geared towards effectively optimizing healthcare resources. As such, we propose precisely and accurately predicting the number of patients with severe COVID-19 prognosis and thus require intense medical intervention. Patients with COVID-19 are typically classified as having severe illness by features such as shortness of breath, low oxygen saturation, and low PaO_2/fraction of inspired oxygen. However, these few features cannot sufficiently distinguish between severe and nonsevere types of patients with COVID-19, as some severe types may lack these or any other symptoms upon admission [3]. Without suitable medical intervention these severe types may progress quickly to a critical condition, resulting in a high risk of mortality [19]. This uncertainty motivates a predictive method that is reliable and efficient, while also making use of alternative features. Early

determination of patient types may enable healthcare professionals to improve their treatment plans and optimize facility resources. This effect demonstrates a need to predict the incoming patients' clinical COVID-19 types. Therefore, our focus lies in the differentiation of severe and non-severe COVID-19 types to support the proper allocation of healthcare staff and space.

While ML techniques have been applied to tackle many aspects of COVID-19, few address the critical question of predicting disease progression upon a patient's admission to the hospital. Some existing studies focused on laboratory/blood-chemistry test results (Gök et al. and Luo et al. [20, 23]). Other studies utilized categorical and binary data extracted from individual patient's electronic health record systems (Hernández-Pereira et al. [22]). However, very few studies included multiple modalities of data, for instance, combination of blood-chemistry and computed tomography (CT) results by Xiong et al. [21], as well as both blood chemistry and clinical data by Chen et al. [3], Jamshidi et al. [25]. Studies focusing on the later Omicron variant included Xu et al. [26] who developed a support vector machine (SVM) classifier for CT images, and Chen et al. [27] who used XGBoost technique based on patients' blood chemistry and clinical features. Other ML techniques for predicting COVID-19 clinical types included random forest (RF) [3,4], as well as multi-algorithm ensemble techniques [4,5].

A major objective of this study is to investigate and evaluate the performances of various ML techniques for COVID-19 severity prediction, as well as to evaluate feature (variable) modalities that provide the most accurate and reliable results. In this study, we will train ML models based on different techniques with patient-level biochemical and clinical feature modalities separately and together as a fusion modality. We will implement logistic regression (LR) with and without regularization, decision tree-based random forest (RF), *k*-nearest neighbors (kNN), and support vector machines (SVM), and compare their predictive power of

severe COVID-19. The inclusion of a variety of state-of-the-art ML techniques allows us to measure and compare different feature modalities' predictive power and interpretability, especially LR and RF that are able to reveal important features to predict severe COVID-19. In addition, we develop these ML models from data collected in both the origin and the Omicron SARS-CoV-2 variants to investigate model robustness across different variants, and generalizability against potential future variants. This study aims to tackle the current challenges of lack of explicit interpretations in many ML applications, to enhance clinical decision support.

A brief overview of each of the ML techniques in this study is provided. LR is a supervised binary classification algorithm which trains a vector of response variables of length equal to the number of features in the data. To make a prediction on a sample, the values of each feature are multiplied with the corresponding coefficient value in the response vector, these products are then summed, and the result is compared to a pre-specified threshold value. If the threshold is exceeded, the sample is given a positive class prediction, otherwise the predicted class is negative. LR is generally sensitive to highly correlated variables, also known as multicollinearity, making predictions less precise [9]. This technique is also rather susceptible to overfitting, in which the model tries to fit too closely on the training set and is less capable of generalizing to unseen prediction sets. A powerful method of reducing overfitting is through regularization or penalization of regression. When LR is penalized (usually using l1 or l2 norms), the multicollinearity issue could be reduced [10]. The best type of regularization employed depends on the problem, thus considered as a hyperparameter in the ML pipeline.

RF is an ensemble learning algorithm that adapts to nonlinearities within the data [11]. Within the "forest", individual decision trees are built on subsets of the training data where the training data are recursively partitioned into "leaves" based on a pre-specified criterion, such as

entropy or Gini impurity. Partitioning ends once each leaf contains samples from only one class (either positive or negative), and the resulting decision tree is a binary classifier. As an ensemble makes a prediction on a sample, each decision tree is making its own prediction, and a majority vote for the predicted class represents the final classification.

kNN classifier assigns samples to their predicted classes with which they share the most similarities, as determined by a chosen distance function [12]. This technique's performance on the choice of *k*, the number of closest "neighbors" to the sample the algorithm consults when predicting a class. The determined distance function distinctly impacts model performance, and can be included in the hyperparameter tuning in the ML pipeline. While kNN can also be used for unsupervised learning tasks, we applied kNN as a supervised classifier for this study.

Lastly, SVMs are classifiers trained to create boundaries between classes in the high dimensional feature space. SVM aims to maximize the distance between samples of different classes. This decision boundary is referred to as a hyperplane, and its geometry allows for application to both linear and non-linear problems. The dimensionality of the problem may be varied based on the choice of the hyperplane; as such, it is also included as a hyperparameter in designing the ML pipeline.

Methods

Data Source

Our study uses two distinct datasets covering different pandemic periods. The first set includes 362 patients diagnosed with COVID-19 upon admission to Wuhan Union Hospital in China from January to March 2020. This set is previously reported on by Chen et al. [3], and serves as a comparison baseline in our study. Among these 362 patients, 148 were determined to

be in severe condition according to guidelines established by the National Health Commision of China and the American Thoracic Society [26, 27], while the remaining 214 were designated non-severe types. Severe types were categorized by meeting at least one the following criteria: (i) respiratory rate >30 breaths per minute; (ii) oxygen saturation <93% at rest; or (iii) PaO/fraction of inspired oxygen <300 mm Hg (40 kPa). Considered to be the original SARS-CoV-2 variant, this set is referred to by the shortening "original" in the rest of this study. The second consists of 1000 patients admitted to Wuhan Union Hospital in China from December 2022 to January 2023, in which patients were diagnosed with the SARS-CoV-2 Omicron variant. Using the same guidelines outlined earlier, 500 of these patients presented with severe stage COVID-19, whereas the other 500 were deemed non-severe. General comparisons of the data set are organized in Table 1. The patients with Omicron variant were age group and gender matched with the patients infected with the original variant. All patients were confirmed to be positive for COVID-19 by two independent quantitative reverse transcriptase-polymerase chain reaction tests before inclusion in this study.

All patients were comprehensively evaluated before admitting to the hospitals. Their fully deidentified, anonymous biomedical data were extracted from the electronic health record system. All participants were informed about the study, agreed to participate, and signed written consent. An institutional review board (IRB) application was submitted and approved by the Wuhan Union Hospital, Tongji College of Medicine, Huazhong University of Science and Technology (IRB approval #IEC-J-345), where the original data were collected. The de-identified patient information comprised two main modalities of biomedical features.The first feature modality had 26 distinct laboratory testing features obtained from blood biochemical tests, most of which were continuous real values of the readings. The specifics of these tests are

detailed by our prior study [3]. We refer to this feature modality as "biochemical" hereinafter. The second is a total of 26 features of one-hot encoded binary values indicating the presence of pre-existing conditions, comorbidities, symptoms, and other common demographic information. This modality was referred to as "clinical features". A complete description of the features across these two modalities was present in the supplementary materials of our prior study [3]. Together, features from these modalities appended into a single corpus of de-identified patient data with 52 multimodal features. This was referred to as the "fusion" set, as it fused across continuous, real-valued biochemical features and binary clinical features. We note that the specific features of respiratory rate, oxygen saturation, and fraction of inspired oxygen were excluded from our predictive feature list as they were the original clinical standard to determine COVID-19 severity.

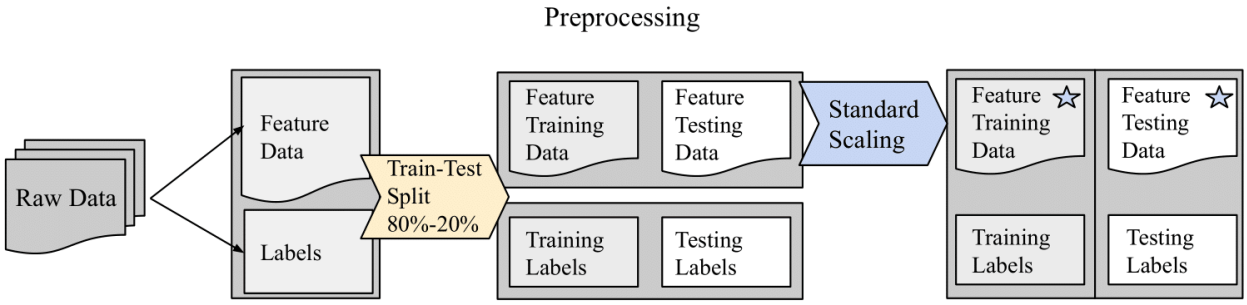
Table 1: Characteristics of COVID-19 Data Sets

Attribute	**Original SARS-CoV-2 Variant**	**Omicron Variant**
Date collected	January-March 2020	December 2022-January 2023
Location collected	Wuhan Union Hospital, China	Wuhan Union Hospital, China
Number of patients (N)	362	1000
Prevalence of severe type	148 (40.9%)	500 (50%)

ML Classification Pipeline Development

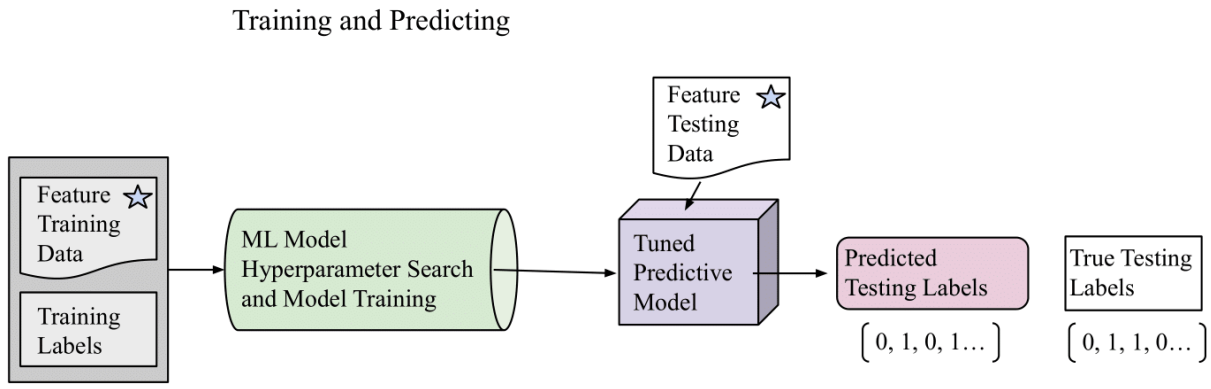
We elected to explore the performance of several competing state-of-the-art machine learning (ML) techniques, including RF, kNN, and support vector machines. To acknowledge LR's popularity in the field [5, 6, 10, 21, 22, 25, 28], we included it as a benchmark classification method in this study as well. All ML techniques were used as supervised binary classifiers.

Utilizing a Google Colaboratory notebook hosting Python 3.10 and a variety of packages from Sci-Kit Learn [13] (see Table 1 in Supplementary Materials for the full list), we developed end-to-end ML frameworks for each of the four ML techniques discussed earlier to predict COVID-19 severity types from patients' clinical, biochemical, and fused feature modalities. Non-severe COVID-19 types were labeled as 0 and severe types were labeled as 1, and each ML technique was constructed to perform a binary classification (predicting 0 or 1) based on different sets of input features. For both the original variant and Omicron data sets, we evaluated ML performance of training and testing on each modality separately and fused. For a given set of data, the corpus was randomly partitioned into training and hold-out testing sets by an 80% to 20% split, respectively. The data was then preprocessed by a standard scaler, in which values were converted to their Z-scores respective of the features. The standard scaler reduces the effect of the variety in ranges that appear in the biochemical and fusion data.



For each ML technique, a grid search method of hyperparameter tuning was utilized to optimize the models' performance via the GridSearchCV package. As previously discussed, the hyperparameters of a ML technique may affect its performance. To account for this, we tuned the following: LR's penalization type, RF's number of trees, tree depth, and criterion, kNN's number of neighbors, distance function and weighting, and SVM's kernel. The resulting optimal hyperparameters for each ML technique were detailed in Table 2 of Supplementary Materials.

Upon the completion of training, the model was applied to predict on the hold-out data. This process was repeated 10 times generating different train-test splits to add to the robustness of the model, as well as avoid overfitting and establish an average performance.



Performance of each model based on the different ML techniques is evaluated based on the model's true positive rate (TPR) (1), true negative rate (TNR) (2), false positive rate (FPR) (3), receiver operating characteristic (ROC) curve, and associated area under the ROC curve (AUC). Models were also evaluated for their variation in performance given the 10 different random training-testing splits. When the model made an individual prediction, it would fall into exactly one of the following categories: a true positive (TP), a false positive (FP), a true negative (TN), or a false negative (FN).

Many of these performance metrics were known by a variety of terms. The TPR is cited as the sensitivity or recall of the model. In essence, this is the probability that the model would predict a positive result (i.e., severe COVID-19 clinical type in this study), conditioned on the patient being truly positive. The TNR was also called the specificity or selectivity of the model. This metric described the probability of a model predicting a negative result (non-severe type) given the patient was truly negative. The FPR is calculated as 1 minus the TNR [7].

$$Sensitivity,\ Recall\text{: } TPR = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity,\ Selectivity\text{: } TNR = \frac{TN}{TN + FP} \quad (2)$$

$$FPR = 1 - TNR = \frac{FP}{FP + TN} \quad (3)$$

The TPR and FPR were utilized when plotting the ROC curve with the FPRs on the x-axis and TPRs on the y-axis. An ideal classifier, theoretically, should have a false positive rate of 0 and a true positive rate of 1, making the "curve" appear as a right line. The associated ideal AUC should be 1, as the axis ranges were between [0,1]. A model with no predictive power would obtain an AUC close to 0.5, equivalent to a purely random prediction. Our choice of model performance comparison based on AUC as opposed to F-measure or accuracy was AUC's proven advantage of being more reliable than the older methods [8].

Feature Importance and Model Interpretability

One of the advantages of certain ML techniques is their interpretability for more informed clinical decision support. Of the methods tested in this investigation, we aimed to glean insights on the data from both the LR and RF models. The resulting feature coefficient vector begot from training LR indicated what the machine "learned" from the data, as the largest coefficient corresponded to the highest importance predicting severe COVID-19 clinical type. Pertaining to RF, feature importance was quantified by a feature's Gini impurity score, which was computed at the time of training. By averaging the feature rankings over the 10 runs, we were able to compare feature importance identified by LR versus RF, as well as different feature importance across the original SARS-CoV-2 variant and the Omicron variant. Feature rankings were also analyzed and compared through the lens of each feature modality independently and fused. These comparisons allowed us to validate our models trained on the original data set, in

that we would cross check our results with other studies, as well as more traditional statistical studies aimed at identifying such features. Upon validation, we offered new analyses to identify critical features with the most predictive power of differentiating severe COVID-19 patients.

Comparison of ML Classification Performance across SARS-CoV-2 Variants, Feature Modalities, and ML Techniques

To compare ML predictive power of COVID-19 severity on the original and Omicron datasets with individual and fused modalities, we opted for the following design. Each data set underwent the pipeline defined earlier, in which the set was used to train and test the model for its predictive power that was measured by various performance measures (e.g., TPR, FPR, AUC, etc). To then evaluate the robustness and generalizability of the developed models, we elected to swap and cross test with the other variant's data. In other words, a model trained on the original COVID-19 variant was tested with both the original and Omicron variant datasets. This analysis was mirrored for models trained on the Omicron variant data as well. For this cross testing, the testing data was standardized according to the scaling scheme fitted from the model's training data. In other words, the scalar fitted to the original patient training data transformed the Omicron patient testing data during the preprocessing step. During cross testing, the entire corpus was used as a hold-out testing set, as the models were never trained with the other variant's data. The cross-set testing was evaluated for its performance using the same metrics as the same-set testing, as we aimed to identify differences between the two variants' data.

Results

Model Performance

Upon running each ML model pipeline for 10 different runs, we calculate the average sensitivity (true positive rate; TPR), specificity (true negative rate; TNR), false positive rate (FPR), and area under curve (AUC). These performance metrics were given by each model with each of the three modalities: biochemical, clinical, and fusion. We validated that the computed average AUC is equivalent to the AUC of the composite or average ROC curve, hence there is no need for their distinction. These values were tabulated according to which dataset was used for training each ML model, with Table 2 being models trained on original SARS-Cov-2 variant and Table 3 being models trained on the Omicron variant.

Among all ML models trained with the original variant across all three modalities (see Table 2), SVM's fusion model gave the highest sensitivity at 78.5%, specifically when tested with the Omicron variant data. Few others could achieve a TPR greater than 70% with exceptions occurring only when testing with Omicron variant data. Specificity rates ranged between 76.3-91.0%, with maximum being achieved by the RF model using fused feature modalities and tested with Omicron data. AUC values fell between 61.1-82.4%, where the maximum was again achieved by the SVM model on fused modalities cross-tested with Omicron variant data.

In general, ML models developed from original SARS-CoV-2 data performed as well or better across all performance metrics when tested with newer Omicron data, compared to testing on the original variant. Regarding TPR, the range of values obtained when testing with original COVID-19 data was 41.7-66.9% while those acquired from testing with Omicron variant data were 53.1-78.5%. Similar trend arised in the comparison of TNR where the original tested rates

were between 76.3-87.0%, while Omicron tested rates were higher at 80.3-91.0%. Lastly, AUC values when testing with original variant data were between 61.1-74.2%, while those from testing with Omicron variant were again higher at 70.5-82.4%.

For the performance across three feature modalities, in most cases the most predictive modality was fusion, i.e., fused features across biochemical and clinical modalities. Fusion performed consistently well across all four ML techniques. With reference to both TPR and AUC, fusion outperformed the biochemical modality in all models trained on the original SARS-CoV-2 variant. This was also observed for most models with respect to TNR, in which the exceptions were minute (within 4% difference). Contrasting with ML models trained on clinical modality alone, models trained on fused features outperformed the same models trained on clinical feature modality only in most metrics across all four ML techniques. Models trained on fusion datasets were especially preferred in instances of testing with the Omicron variant data, while models developed from clinical feature modality only were the best performers when testing with original SARS-CoV-2 variant data. The only two occasions where the ML models developed from clinical modality alone achieved higher AUC scores than the fusion models were in the case of RF tested with original variant data, and kNN tested with original variant data.

Table 2: Mean Original Model Performance (10 runs)

Regularized Logistic Regression (LR): Trained on Original Variant					
Test Set	Modality	TPR(%)	TNR(%)	FPR(%)	AUC
Original	Biochemical	53.5%	82.8%	17.2%	0.682
	Clinical	64.6%	77.9%	22.1%	0.713
	Fusion	66.9%	81.5%	18.5%	0.742
Omicron	Biochemical	66.8%	80.3%	19.7%	0.735

	Clinical	70.6%	82.0%	18.0%	0.763
	Fusion	75.8%	81.2%	18.8%	0.785
Random Forest (RF): Trained on Original Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Original	Biochemical	53.0%	83.7%	16.3%	0.683
	Clinical	65.9%	76.3%	23.7%	0.711
	Fusion	57.3%	79.9%	20.1%	0.686
Omicron	Biochemical	68.0%	90.3%	9.7%	0.791
	Clinical	67.6%	81.7%	18.3%	0.747
	Fusion	70.5%	**91.0%**	**9.0%**	0.808
k-Nearest Neighbors (kNN): Trained on Original Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Original	Biochemical	41.7%	80.4%	19.6%	0.611
	Clinical	55.0%	87.0%	13.0%	0.710
	Fusion	56.2%	83.8%	16.2%	0.700
Omicron	Biochemical	53.1%	88.0%	12.0%	0.705
	Clinical	60.1%	84.2%	15.8%	0.722
	Fusion	70.7%	90.2%	9.8%	0.805
Support Vector Machine (SVM): Trained on Original Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Original	Biochemical	50.5%	86.6%	13.4%	0.685
	Clinical	64.6%	79.7%	20.3%	0.721
	Fusion	63.4%	82.6%	17.4%	0.730
Omicron	Biochemical	63.1%	82.6%	17.4%	0.728

	Clinical	71.1%	83.6%	16.4%	0.774
	Fusion	**78.5%**	86.3%	13.7%	**0.824**

Within the techniques trained with Omicron variant data (Table 3), the highest sensitivity achieved was 92.4% by SVM fusion model when tested with Omicron variant data. Following closely with a TPR of 92.0% was the LR fusion model tested on Omicron variant data. The specificity rates were in the range 73.5-90.8% with the highest rate achieved by the LR fusion model tested with Omicron variant data. Overall AUC values ranged between 66.8 and 91.4%, where the LR fusion model was the top performer.

Similarly to what we found when analyzing the performance of ML models trained on original variant data, models trained with Omicron variant data performed better when tested with Omicron variant data compared to testing with the original variant across all performance metrics (Table 3). Sensitivity ranged 76.6-92.4% when testing with Omicron variant data. When testing with original variant data, sensitivity ranged 53.9-78.4%. Specificity showed a similar pattern, as the Omicron-tested values were between 77.3-90.8%, while original variant tested values were lower at 73.5-82.9%. Also following this pattern, AUC values of the Omicron tested set were between 77.7-91.4%, whilst those from the original variant tested were 66.8-76.9%.

As for the performance across feature modalities, Table 3 also demonstrated models with fusion features as the overall best performers. Models with fused feature modalities outperformed models with biochemical modality alone in every performance metric among ML models trained on Omicron variant data. Performances were similar between models with the fusion and clinical feature modalities, with fusion achieving slightly higher values in the majority of scenarios. We again only observed better AUC values in LR and SVM with clinical feature

modality alone than in counterpart models with fusion features when tested on original variant data.

Table 3: Mean Omicron Model Performance (10 runs)

Regularized Logistic Regression (LR): Trained on Omicron Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Omicron	Biochemical	84.1%	88.3%	11.7%	0.862
	Clinical	82.3%	82.7%	17.3%	0.825
	Fusion	92.0%	**90.8%**	**9.2%**	**0.914**
Original	Biochemical	58.3%	77.9%	22.1%	0.681
	Clinical	77.4%	76.4%	23.6%	0.769
	Fusion	71.1%	79.3%	20.7%	0.752
Random Forest (RF): Trained on Omicron Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Omicron	Biochemical	87.5%	87.7%	12.3%	0.876
	Clinical	80.3%	77.3%	22.7%	0.788
	Fusion	89.5%	89.8%	10.2%	0.896
Original	Biochemical	71.6%	76.1%	23.9%	0.738
	Clinical	78.4%	73.5%	26.5%	0.759
	Fusion	76.5%	76.9%	23.1%	0.767
k-Nearest Neighbors (kNN): Trained on Omicron Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Omicron	Biochemical	75.7%	86.0%	14.0%	0.809
	Clinical	76.6%	78.8%	21.2%	0.777
	Fusion	83.2%	88.2%	11.8%	0.857

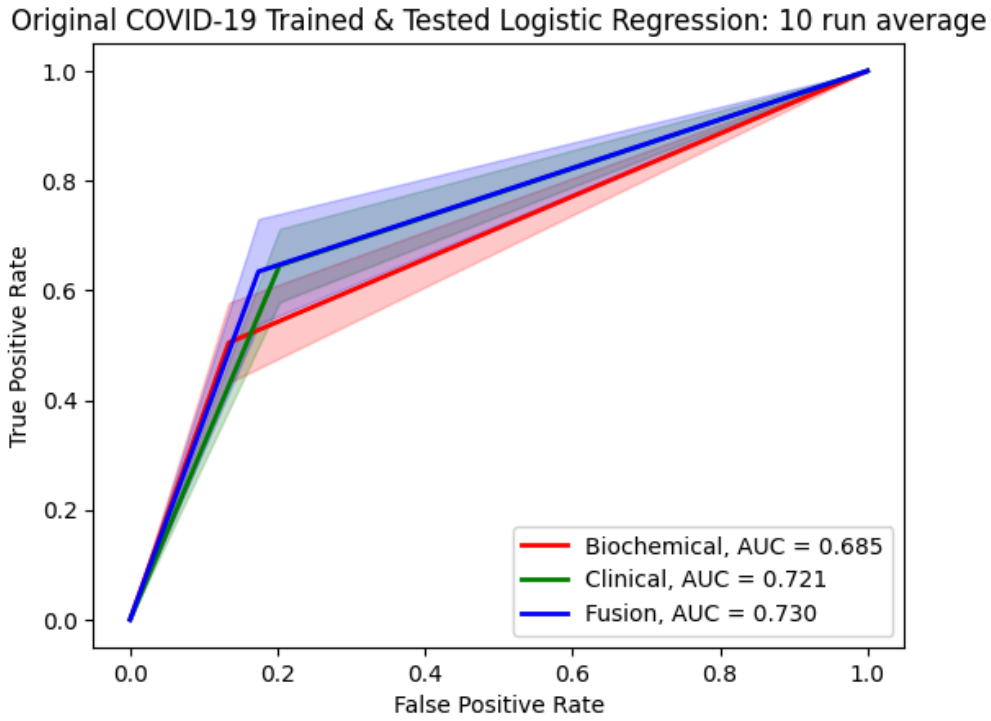
Original	Biochemical	53.9%	79.8%	20.2%	0.668
	Clinical	69.1%	78.0%	22.0%	0.735
	Fusion	64.5%	82.9%	17.1%	0.737
Support Vector Machine (SVM): Trained on Omicron Variant					
Test Set	Modality	TPR	TNR	FPR	AUC
Omicron	Biochemical	82.1%	87.9%	12.1%	0.850
	Clinical	80.6%	82.0%	18.0%	0.813
	Fusion	**92.4%**	88.5%	11.5%	0.905
Original	Biochemical	57.4%	77.9%	22.1%	0.677
	Clinical	76.3%	76.7%	23.3%	0.765
	Fusion	70.7%	78.4%	21.6%	0.746

To visualize and compare across the three feature modalities, we constructed the following receiver operating characteristic (ROC) plots. For each of the four ML techniques, ROC plots were generated for the same-variant and cross-variant testings with a total of 4 combinations. All 16 plots were presented in the Supplemental Materials. For brevity's sake, we demonstrated four graphs from LR in Figures 2(a)-(d). Each graph showed the composite ROC obtained by averaging sensitivity (TPR) and 1-specificity (FPR) over 10 runs for each modality. The shaded regions denote one standard deviation from the mean of the TPR. As depicted in the legend, the red, green, and blue lines represented the performance of models with biochemical, clinical, and fusion feature modalities, respectively.

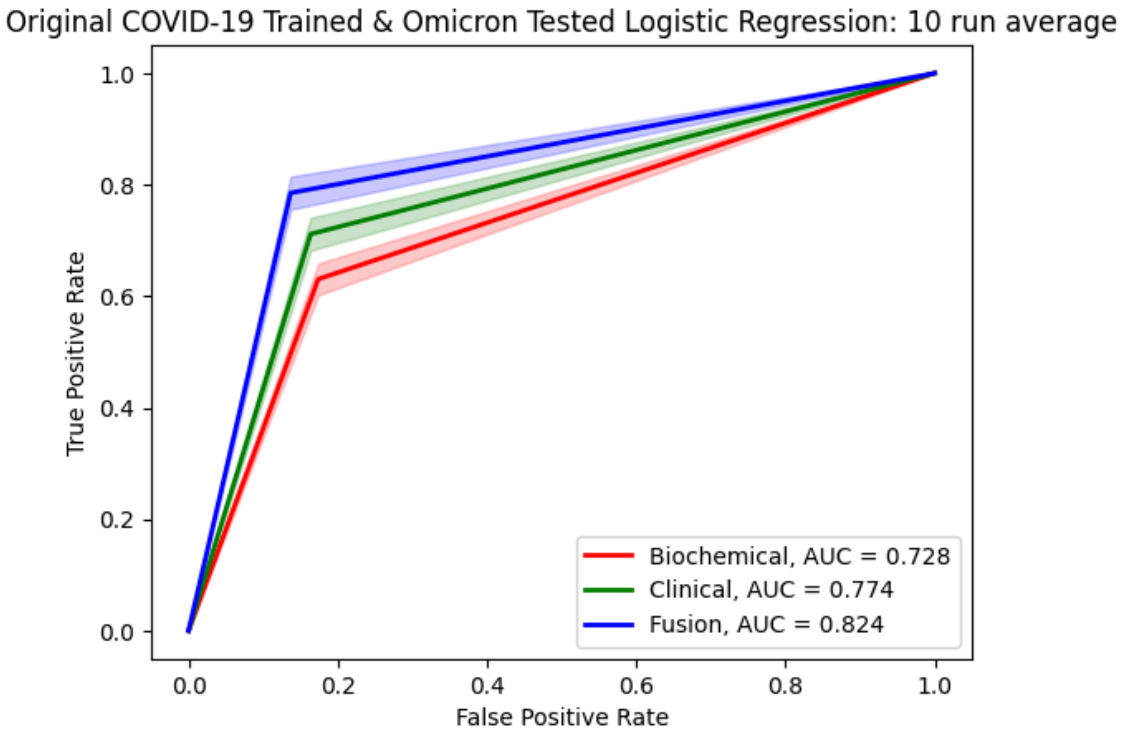
In general, models with fusion modality performed equally well and sometimes better than models with either a single biochemical or clinical feature modality, as shown in Tables 2-3. This was also confirmed by the ROC curves and area under curve (AUC) values, where models

with fusion feature modality were consistently above other models in both training and testing combinations. In summary, these results illuminated a unique and previously unknown attribute of ML models trained in original SARS-CoV-2 variant but tested with later Omicron variant, and there was a significant variation in performance across 10 runs. This was not observed in three other same and cross-variant training-testing combinations. It was also worth noting that this pattern was also observed in RF, kNN, and SVM ML techniques, and was not unique to LR.

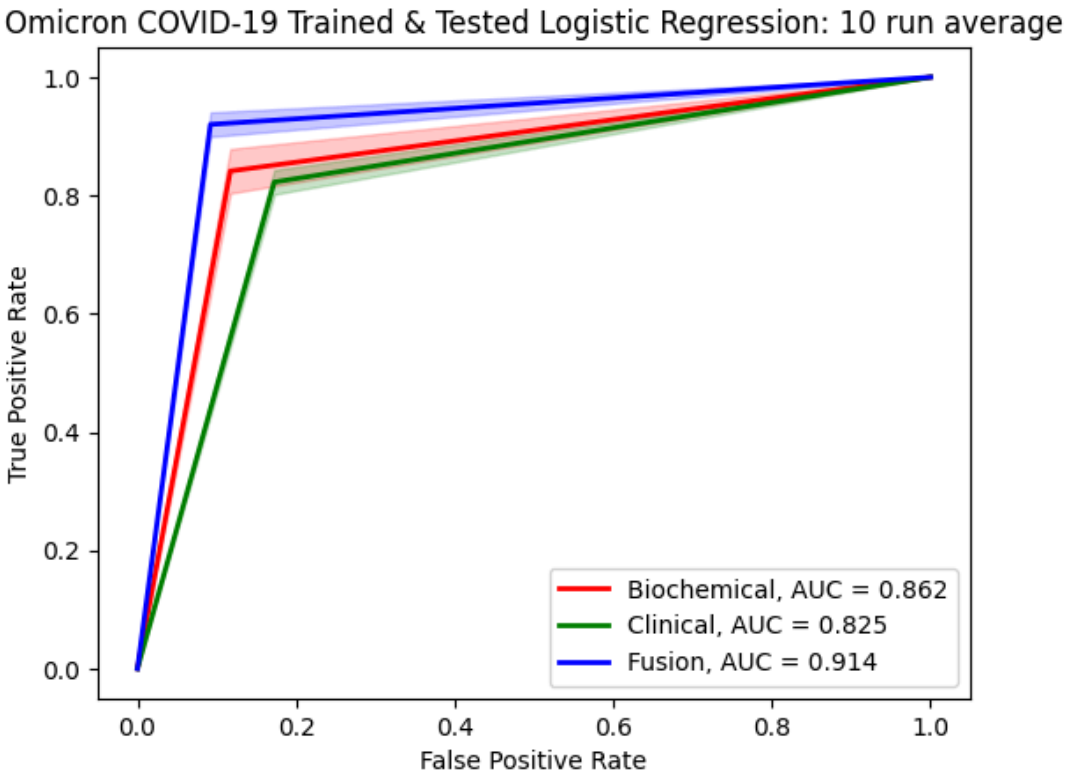
Figures 2(a)-(d)



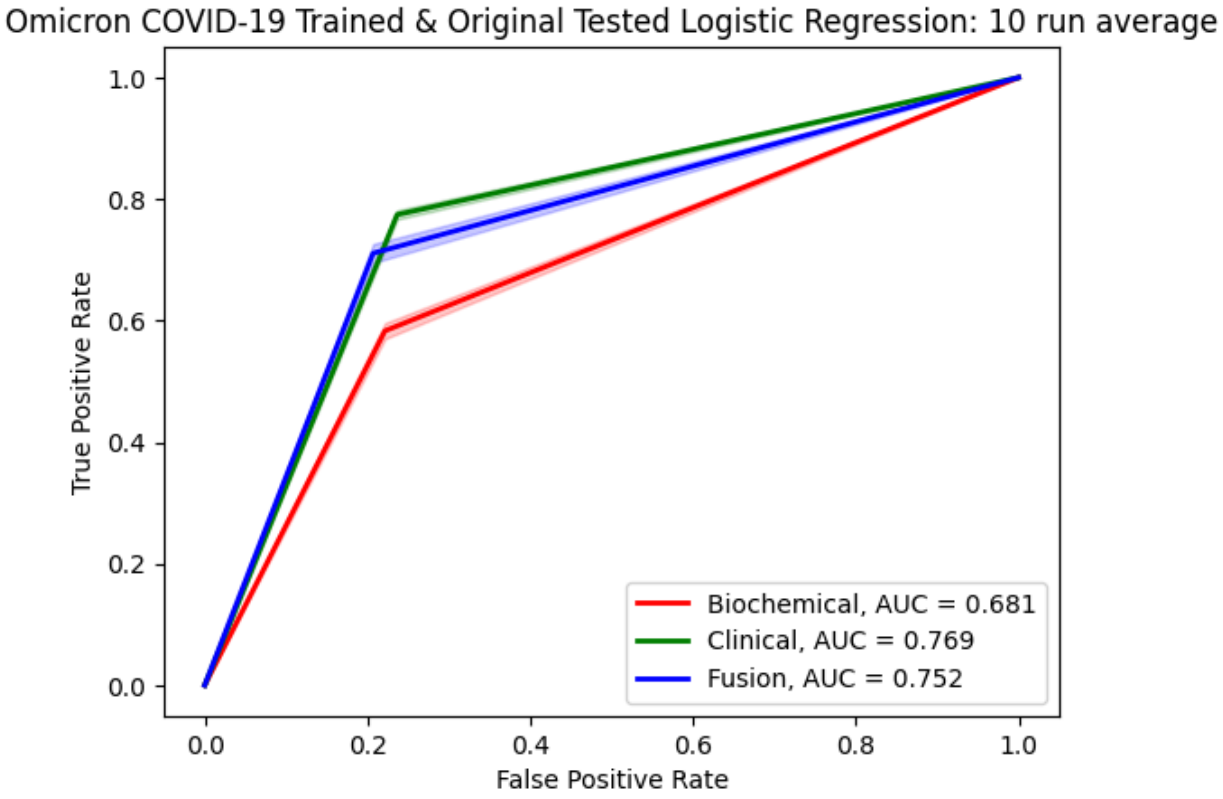
2(a)



2(b)



2(c)



2(d)

Feature Importance Ranking for Clinical Decision Support on COVID-19 Severity

During each run of ML models, the feature coefficient vectors from the tuned LR model and Gini impurity-based feature importances from RF were recorded for feature ranking. At the end of the 10 runs, these weights and importances were averaged with respect to the specific training data set and feature modality. This resulted in an overall ranking of features of each modality separately and fused when trained on either SARS-CoV-2 variant, according to LR and RF. LR's associated coefficient vector would be real valued, while RF's Gini-importance were probabilities in [0, 1].

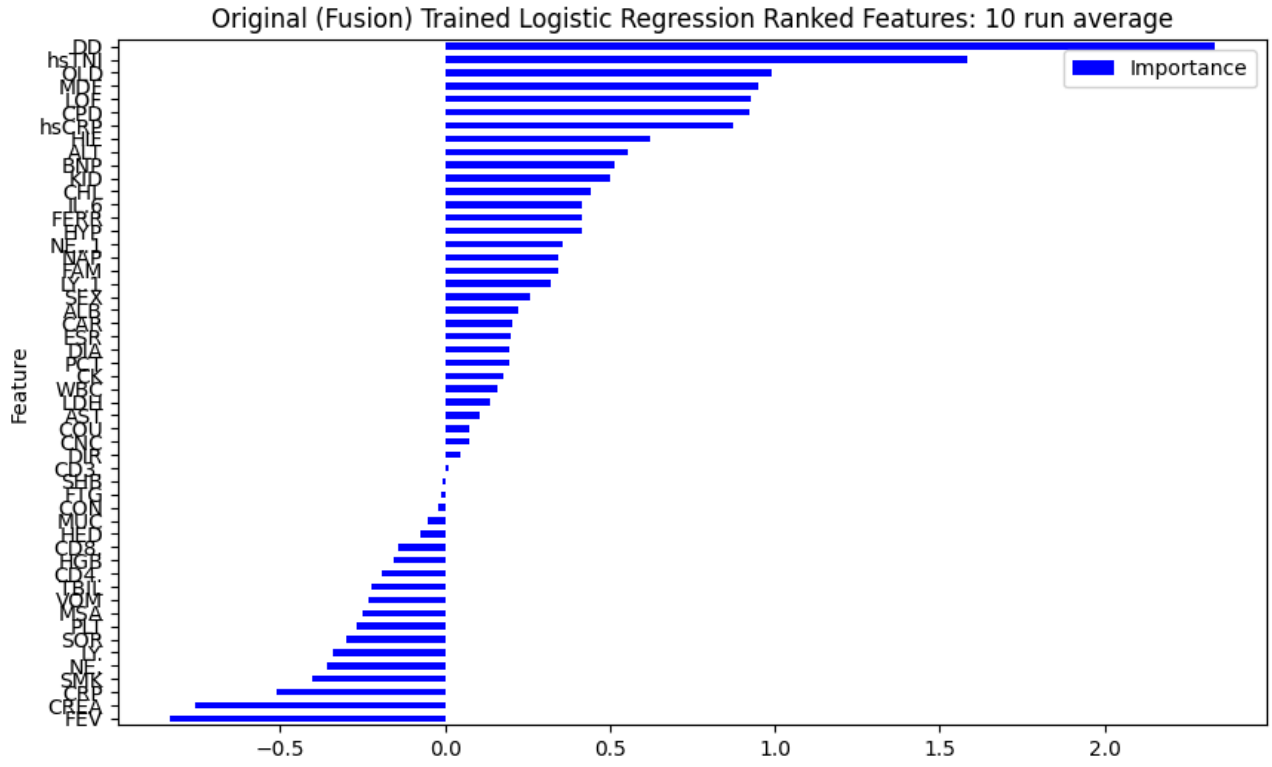
ML performance results showed that fusion features of both biochemical and clinical modalities generally yielded the most reliable predictions in either original or Omicron variant

data. Fusion across the two modalities also covered features more comprehensively. Therefore, we focused on fusion features to demonstrate feature importance in differentiating severe and non-severe COVID-19 types. These rankings are presented in Figures 3(a)-(d). Feature rankings for each modality separately can be found in the Supplementary Materials. Regardless of ML model techniques or SARS-CoV-2 variants, certain features such as DD (d-dimer; biochemical modality), hsTNI (high sensitivity Troponin I; biochemical), and OLD (age >55 years; clinical modality) were consistently ranked in the top five most predictive features for COVID-19 severity. Features that often appeared in the top ten most predictive features include hsCRP (high sensitivity C-reactive protein; biochemical) and HYP (hypertension; clinical).

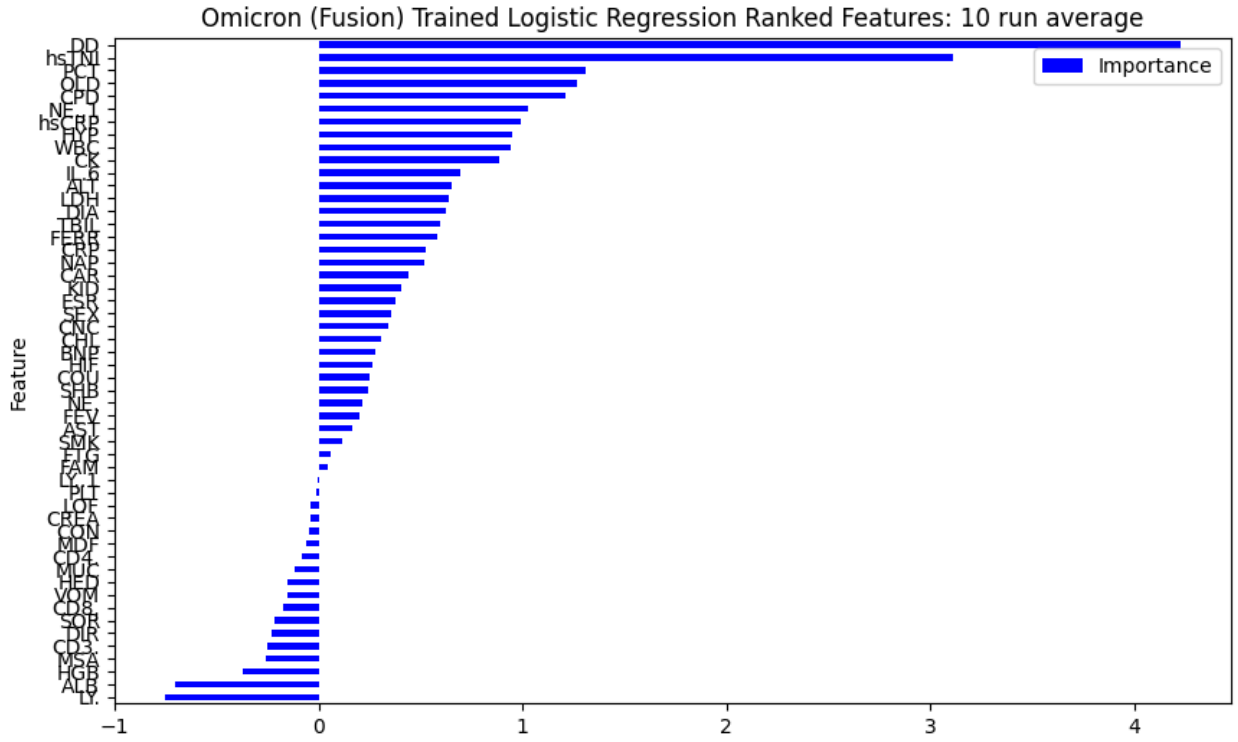
Comparing LR and RFs' respective feature rankings, there were a few differences to note. Primarily, both LR models (figures 3(a),(b)) ranked CPD (chronic obstructive pulmonary disease; clinical) the 6th most predictive feature according to the ML model trained on the original variant; and 5th trained on the Omicron variant data. This result was not mirrored for RF. Instead, we found that both RF models trained on original and Omicron variants (figures 3(c),(d)) identified LY (lymphocyte; biochemical), FERR (ferritin; biochemical), and IL-6 (interleukin-6; biochemical) as important features. Interestingly, the RF models trained on different variants' datasets agreed more on feature ranking than LR trained on different datasets. For example, only LR trained on the original variant identified MDF (mid-grade fever; clinical), LOF (low-grade fever; clinical), and HIF (high-grade fever; clinical) among the top 10 most predictive features, while its counterpart trained on Omicron variant identified PCT (procalcitonin; biochemical), NE.1 (percent of neutrophil; biochemical) and WBC (white blood cell; biochemical) as the most predictive. Such discrepancies in feature rankings were not observed in results from RF models trained on different variants' datasets.

Lastly, there were substantial differences in the range of feature importance when comparing models trained on different variants' datasets. LR's feature coefficients ranged between approximately [-0.83, 2.30] in the original variant, whereas this range was [-0.75, 4.20] for Omicron. Gini impurities obtained from the RF model trained on the original variant dataset were around [0, 0.09], while Gini impurities from RF models trained on Omicron dataset were broader [0, 0.12].

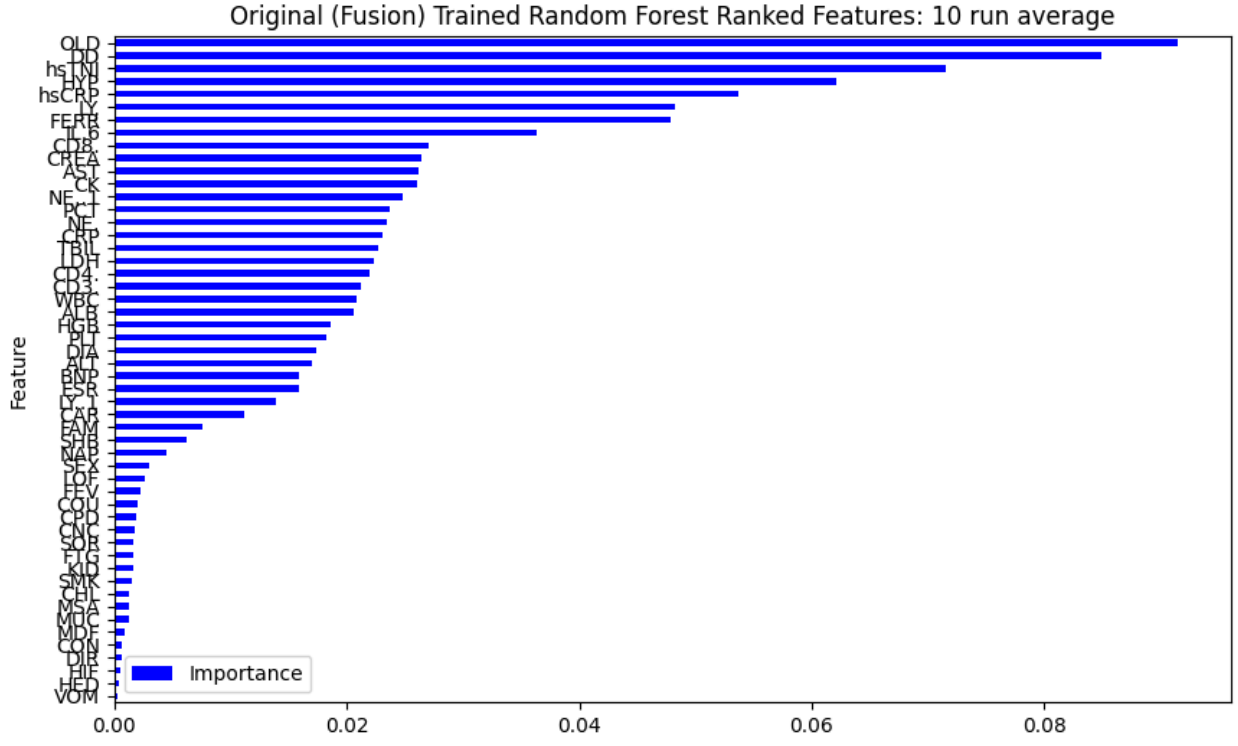
Figures 3(a)-(d)



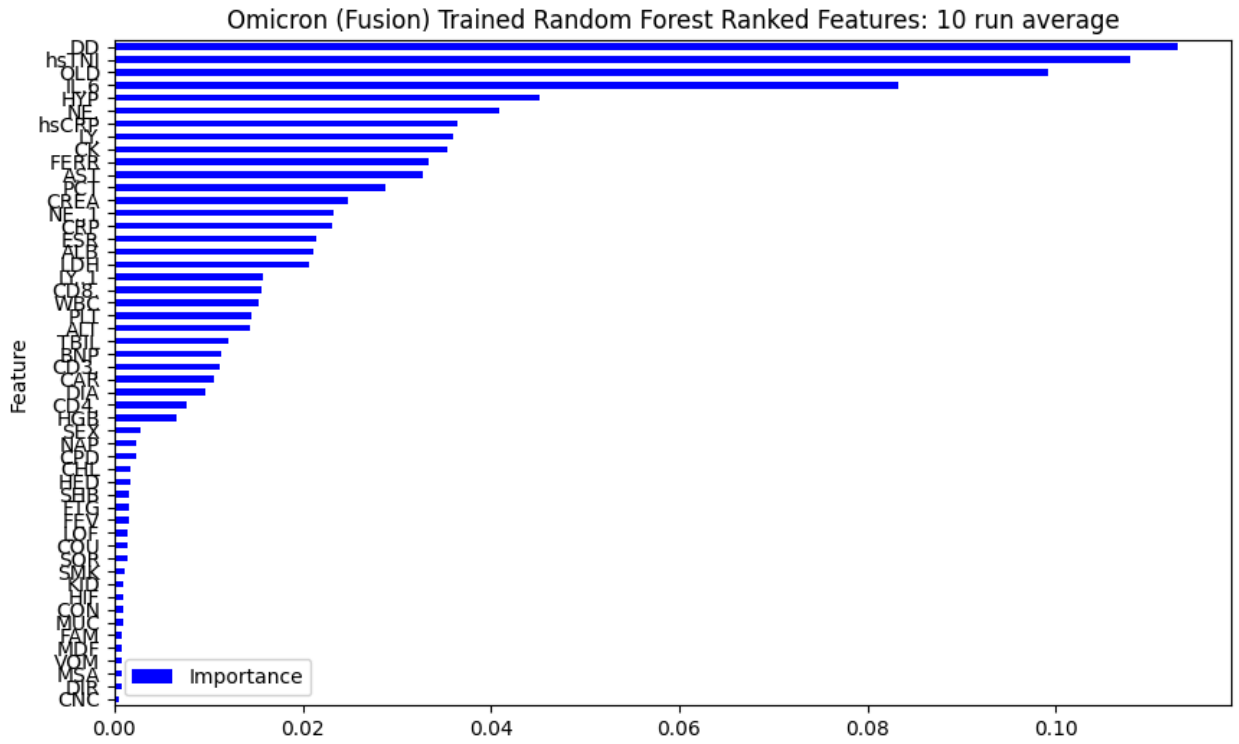
3(a)



3(b)



3(c)



3(d)

Discussion

Main Findings

In this study, we evaluated the predictive power of multiple ML techniques when utilizing two different feature modalities. Not only were we able to compare the predictive capabilities of COVID-19 severity across these modalities, we also discovered the differences of model performance across different SARS-CoV-2 variants. Overall, we found ML to be a powerful tool for predicting COVID-19 severity based on comprehensive individual patient-level data. More importantly, we discovered that fusion of the biochemical and clinical modalities enhanced the predictive power of all types of ML models evaluated in this study, including LR, RF, kNN, and SVM . Models trained on multiple feature modalities have yielded the best

performance in nearly every performance metrics across training and testing sets. These multimodal features are accessible by healthcare systems especially with wide adoption of electronic health record systems. Results can be obtained efficiently from these systems, allowing the predictive ML model as a fast and reliable clinical decision support tool to quickly and effectively identify patients with high risk of severe COVID-19.

The similarity of performance results between the four ML techniques evaluated in this study suggest that the specific choice of modeling technique is less important for the task of predicting severe COVID-19 patients. In general, LR, RF, and SVMs tied as the top performers with their highest AUC scores being 0.914, 0.896 and 0.905, respectively. kNN is the relative lowest, with its highest AUC being 0.857. If model interpretability is important in the clinical decision support of these ML models, then LR and RF should be considered. LR offers the analyst information on which features are positively and negatively associated with the risk of severe COVID-19 However, LR is susceptible to multicollinearity between different features [9]. RF, on the other hand, is more resilient to the multicollinearity issue in the input data [11]. RF model's reliability and robustness is shown in the findings of this study, and further supported by other studies, including Chen et al. [3], Xiong et al. [21], and more.

The feature rankings provided by the two ML models are important for clinical decision making and serve as clues to COVID-19 pathology. Our study indicates that elevated biomarkers such as D-dimer for coagulation and hsTNI and hsCRP as indicators for heart damage are strongly associated with severe COVID-19. Other works have also shown that cardiovascular injury due to COVID-19 is highly associated with adverse patient outcomes [29]. Higher D-dimer is associated with higher risk of progressing to severe stage. Our findings also suggest that patients' clinical information such as being 55 years or older, or having pre-existing

conditions such as hypertension and COPD, could significantly increase the risk of progressing to severe COVID-19. Other studies also confirm age and hypertension as major risk factors for severe COVID-19 [30, 31].

When comparing important features between patients infected by original and Omicron variants, we have identified an increase in the feature weight vector and Gini impurity values, which have not been reported before. This finding suggests that COVID-19 severity became more predictable in more recent variants. We speculate that patient-level data may have higher quality in the Omicron wave than the original variant. This might also explain the higher variability and lower performance in models trained and tested on the original SARS-CoV-2 patient data.

Limitations

One of the hindrances to the generalizability of this framework is the lack of variation in data samples. All patient data were taken at the individual's time of admission to one healthcare facility, the Wuhan Union Hospital, and the majority of patients were of the Han Chinese ethnicity group. This could result in potential sampling bias and the results should be further validated with larger scale mutli-center studies. For this framework to be more robust, incorporation of larger datasets across more demographic groups would be necessary.

Another limitation is that we were not able to evaluate ML model predictability for other major SARS-CoV-2 variants, such as Alpha and Delta. Retrospective studies are needed to comprehensively evaluate the robustness of the developed ML models across different phases of COVID-19 with different dominant variants and subvariants.

Future Directions

There are a plethora of directions for which this study could further investigate. Notwithstanding improvements made on the limitations previously discussed, we acknowledge the existence of other emerging ML techniques that could be explored and evaluated in future work. Given the tentative promise of LR as a predictive tool for COVID-19 clinical decision support, other regression techniques such as the Lasso or Ridge regressions may be useful.

Further explorations shed light on the predictive power of ML techniques from individual-level data collected from patients with other respiratory illnesses. This is especially useful to healthcare systems inundated with patients infected with influenza or respiratory syncytial virus (RSV), to name a few. Using similar ML techniques and leveraging the power of transfer learning, our developed ML pipeline can be further applied to other diseases with similar underlying datasets (e.g., clinical and biochemical). Connecting back with the goal of aiding healthcare resource optimization, a potential application of this work is to simulate burdens on the health system of an unexpected inflow of patients, some of whom are severe patients and therefore need intensive care.

Another future direction to this work would be the incorporation of more data modalities, such as patient-level medical imaging (including X Ray and CT scans) and multi-omics data. Due to the higher dimensionality of imaging modality in relation to the biochemical and clinical modalities, more advanced ML techniques such as a deep convolutional neural network (CNN) need to be applied to match dimensionalities across different modalities.

Acknowledgements

This study is supported by the U.S. Centers for Disease Control and Prevention grant U01CK000677 Building Mathematical Modeling Workforce Capacity to Support Infectious Disease and Healthcare Research (HIRe Modeling Fellowship).

Conflicts of Interest

None declared.

GitHub Repository

A copy of the Google Colaboratory notebook created for this study have been made available publicly through the GitHub repository here:

https://github.com/hnwestpage/Fusion-ML-COVID-19

References

1. Duggal, A., & Mathews, K. S. (2022). Impact of ICU strain on outcomes. Current opinion in critical care, 28(6), 667–673. https://doi.org/10.1097/MCC.0000000000000993

2. COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment Guidelines. National Institutes of Health. Available at https://www.covid19treatmentguidelines.nih.gov/. Accessed [12/06/2023].

3. Chen Y, Ouyang L, Bao FS, Li Q, Han L, Zhang H, Zhu B, Ge Y, Robinson P, Xu M, Liu J, Chen S. A Multimodality Machine Learning Approach to Differentiate Severe and Nonsevere COVID-19: Model Development and Validation. J Med Internet Res 2021;23(4):e23948. doi: 10.2196/23948

4. Patterson BK, Guevara-Coto J, Yogendra R, Francisco EB, Long E, Pise A, Rodrigues H, Parikh P, Mora J and Mora-Rodríguez RA (2021) Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning. Front. Immunol. 12:700782. doi:10.3389/fimmu.2021.700782

5. Shakhovska, N., Yakovyna, V., & Chopyak, V. (2022). A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system. Mathematical biosciences and engineering : MBE, 19(6), 6102–6123. https://doi.org/10.3934/mbe.2022285

6. Cabitza, Federico, Campagner, Andrea, Ferrari, Davide, Di Resta, Chiara, Ceriotti, Daniele, Sabetta, Eleonora, Colombini, Alessandra, De Vecchi, Elena, Banfi, Giuseppe, Locatelli, Massimo and Carobene, Anna. "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests" *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 59, no. 2, 2021, pp. 421-431. https://doi.org/10.1515/cclm-2020-1294

7. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

8. Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 299-310, March 2005, doi: 10.1109/TKDE.2005.50

9. Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. Perspectives in clinical research, 8(3), 148–151. https://doi.org/10.4103/picr.PICR_87_17

10. Rymarczyk T, Kozłowski E, Kłosowski G, Niderla K. Logistic Regression for Machine Learning in Process Tomography. Sensors. 2019; 19(15):3400. https://doi.org/10.3390/s19153400

11. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. https://doi.org/10.1177/1536867X20909688

12. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med 2016;4(11):218. doi: 10.21037/atm.2016.03.37

13. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

14. Janke, A. T., Mei, H., Rothenberg, C., Becher, R. D., Lin, Z., & Venkatesh, A. K. (2021). Analysis of Hospital Resource Availability and COVID-19 Mortality Across the United States. Journal of hospital medicine, 16(4), 211–214. https://doi.org/10.12788/jhm.3539

15. Cabitza, F., Campagner, A., Ferrari, D., Di Resta, C., Ceriotti, D., Sabetta, E., Colombini, A., De Vecchi, E., Banfi, G., Locatelli, M. & Carobene, A. (2021). Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clinical Chemistry and Laboratory Medicine (CCLM), 59(2), 421-431. https://doi.org/10.1515/cclm-2020-1294

16. Cassaro F, Pires L (2020) Can we predict the occurrence of covid-19 cases? considerations using a simple model of growth. Sci Total Environ 728:138834

17. Niazkar M, Niazkar H (2020) Covid-19 outbreak:application of multi-gene genetic programming to country-based prediction models. Electron J Gen Med 17(5):247

18. COVID-19 Epidemiological Update – 16 February 2024. World Health Organization. 2024 Feb 16. URL: https://www.who.int/publications/m/item/covid-19-epidemiological-update-16-february-2024 [accessed 2024-02-20]

19. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323(13):1239–1242. doi:10.1001/jama.2020.2648

20. Gök, E. C., & Olgun, M. O. (2021). Smote-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. Neural Computing and Applications, 33(22), 15693–15707. https://doi.org/10.1007/s00521-021-06189-y

21. Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., & Huang, L. (2022a). Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious Diseases of Poverty*, *11*(1). https://doi.org/10.1186/s40249-022-00946-4

22. Hernández-Pereira, E., Fontenla-Romero, O., Bolón-Canedo, V., Cancela-Barizo, B., Guijarro-Berdiñas, B., & Alonso-Betanzos, A. (2021). Machine learning techniques to predict different levels of hospital care of covid-19. Applied Intelligence, 52(6), 6413–6431. https://doi.org/10.1007/s10489-021-02743-2

23. Luo J, Zhou L, Feng Y, Li B, Guo S (2021). The selection of indicators from initial blood routine test results to improve the accuracy of early prediction of COVID-19 severity. PLoS ONE 16(6):e0253329. https://doi.org/10.1371/journal.pone.0253329

24. Rekha K, Shailja S, Taha A, Mohammad A K, Nahed A. El-S, Firzan N, Perumal A D, Kuldeep D. Emergence of SARS-CoV-2 Omicron (B.1.1.529) variant, salient features, high global health concerns and strategies to counter it amid ongoing COVID-19 pandemic. Environmental Research. Volume 209, 2022, 112816, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2022.112816. (https://www.sciencedirect.com/science/article/pii/S0013935122001438)

25. Jamshidi, E., Asgary, A., Tavakoli, N., Zali, A., Setareh, S., Esmaily, H., Jamaldini, S. H., Daaee, A., Babajani, A., Sendani Kashi, M. A., Jamshidi, M., Jamal Rahi, S., & Mansouri, N. (2022). Using Machine Learning to Predict Mortality for COVID-19 Patients on Day 0 in the ICU. Frontiers in digital health, 3, 681608. https://doi.org/10.3389/fdgth.2021.681608

26. Novel coronavirus pneumonia diagnosis and treatment plan (provisional 7th edition). Webpage in Chinese. National Health Commission of China. 2020 Mar 04. URL: http://www.nhc.gov.cn/yzygj/s7652m/202003/a31191442e29474b98bfed5579d5af95.shtml

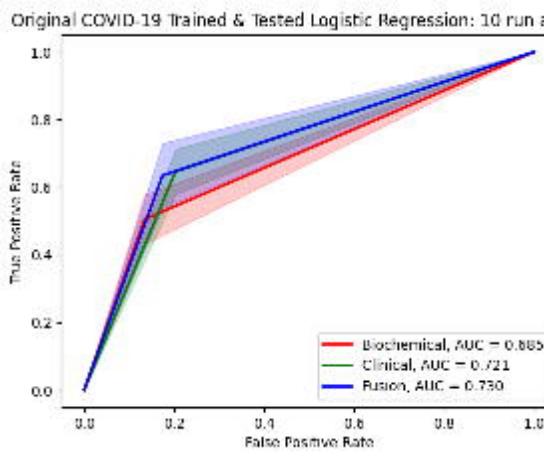
27. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. Am J Respir Crit Care Med 2019 Oct 01;200(7):e45-e67. [doi:10.1164/rccm.201908-1581st]

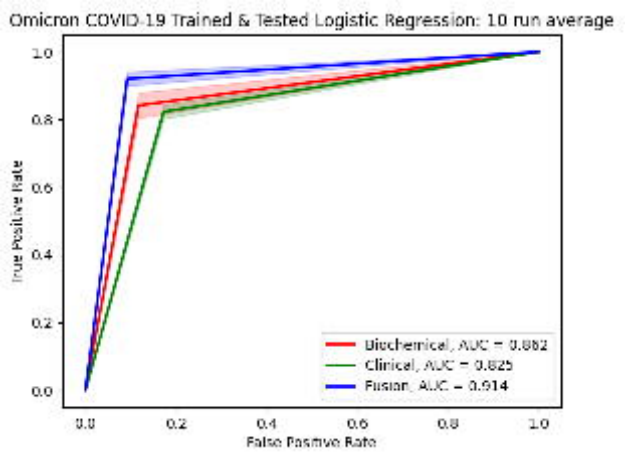
28. Saegerman C, Gilbert A, Donneau A-F, Gangolf M, Diep AN, Meex C, et al. (2021) Clinical decision support tool for diagnosis of COVID-19 in hospitals. PLoS ONE 16(3): e0247773. https://doi.org/10.1371/journal.pone.0247773

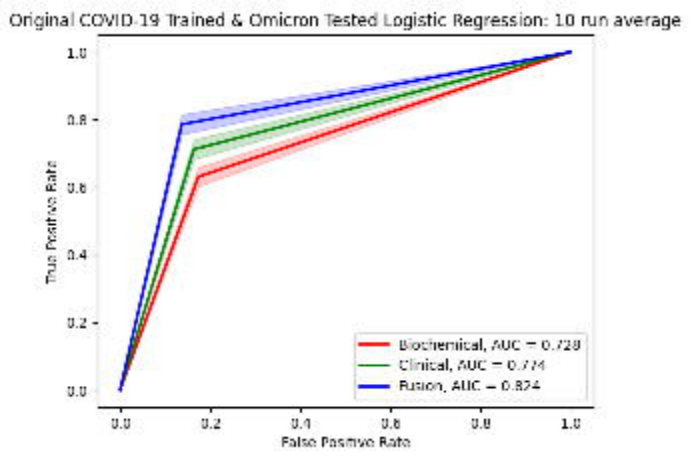
29. Shi S, Qin M, Shen B, et al. Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China. JAMA Cardiol. 2020;5(7):802–810. doi:10.1001/jamacardio.2020.0950

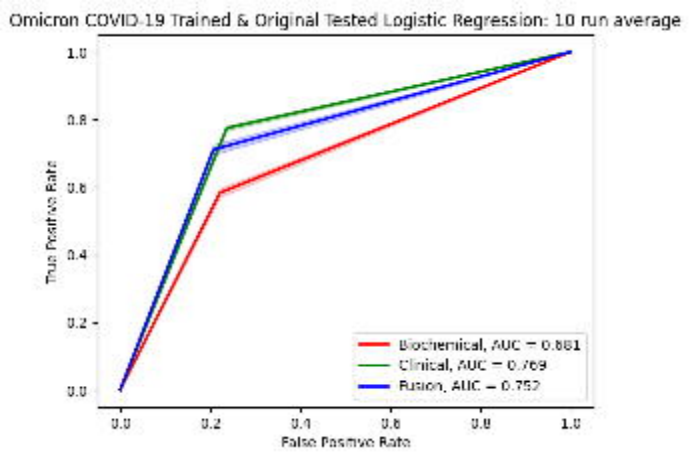
30. Almazeedi, S., Al-Youha, S., Jamal, M. H., Al-Haddad, M., Al-Muhaini, A., Al-Ghimlas, F., & Al-Sabah, S. (2020). Characteristics, risk factors and outcomes among the first consecutive 1096 patients diagnosed with COVID-19 in Kuwait. EClinicalMedicine, 24, 100448. https://doi.org/10.1016/j.eclinm.2020.100448

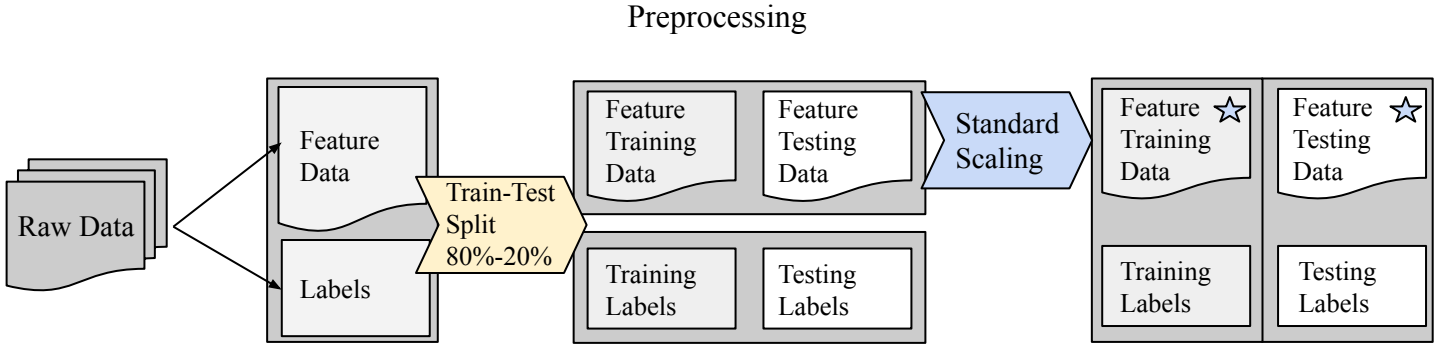
31. Suleyman G, Fadel RA, Malette KM, et al. Clinical Characteristics and Morbidity Associated With Coronavirus Disease 2019 in a Series of Patients in Metropolitan Detroit. JAMA Netw Open. 2020;3(6):e2012270. doi:10.1001/jamanetworkopen.2020.12270

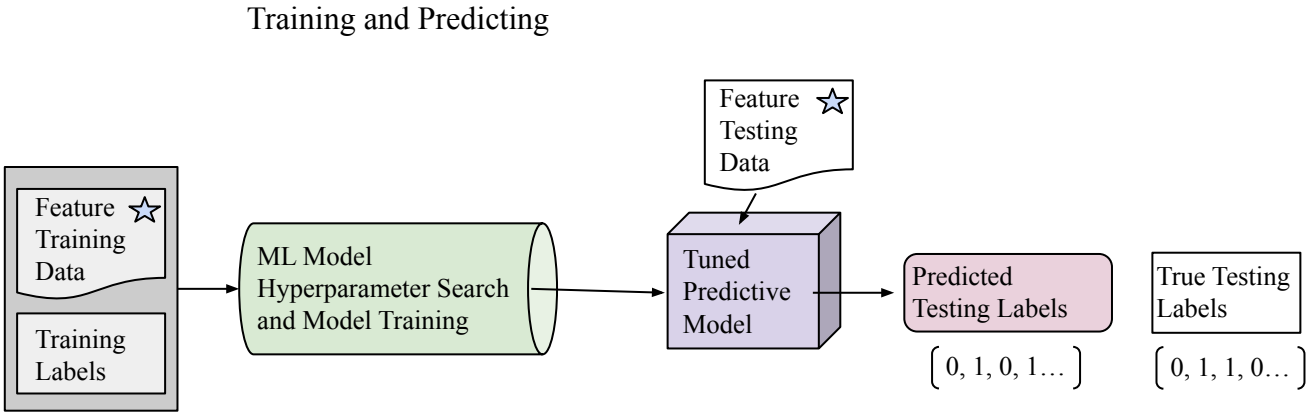

Original COVID-19 Trained & Tested Logistic Regression: 10 run average
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Biochemical, AUC = 0.885
Clinical, AUC = 0.721
Fusion, AUC = 0.730

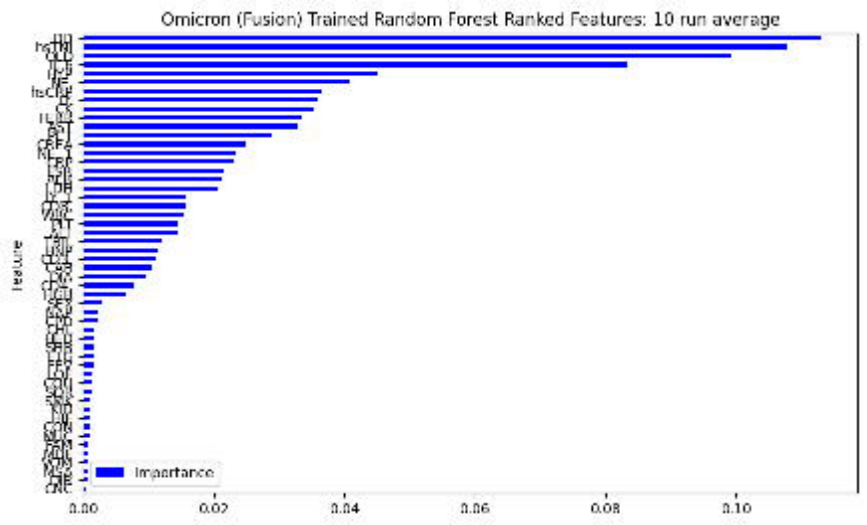

Omicron COVID-19 Trained & Tested Logistic Regression: 10 run average
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Biochemical, AUC = 0.882
Clinical, AUC = 0.825
Fusion, AUC = 0.914

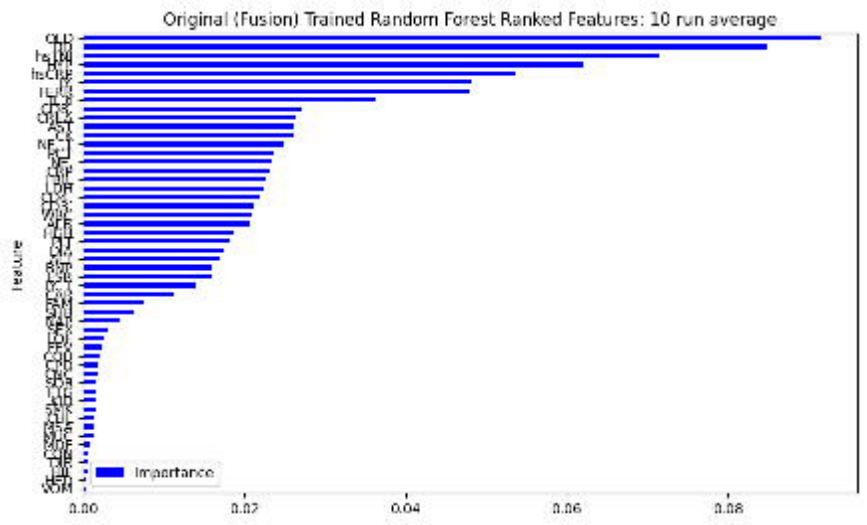

Original COVID-19 Trained & Omicron Tested Logistic Regression: 10 run average
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Biochemical, AUC = 0.728
Clinical, AUC = 0.774
Fusion, AUC = 0.824

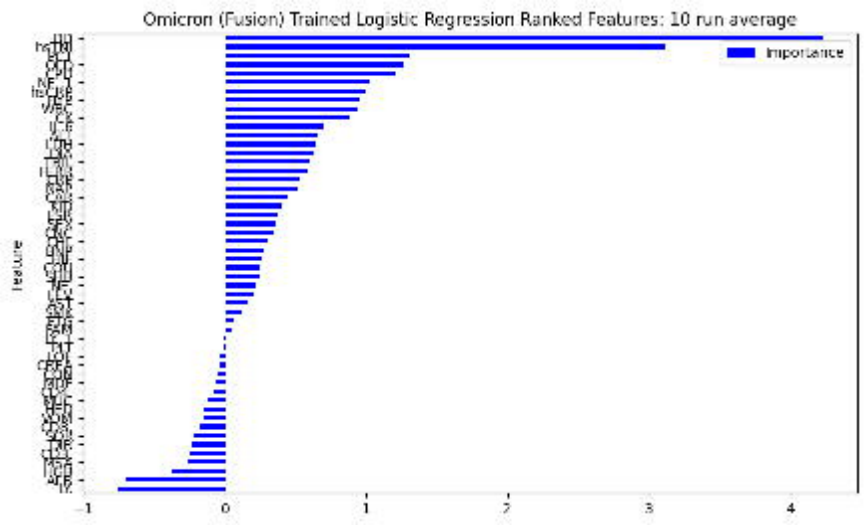

Omicron COVID-19 Trained & Original Tested Logistic Regression: 10 run average
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Biochemical, AUC = 0.881
Clinical, AUC = 0.769
Fusion, AUC = 0.752

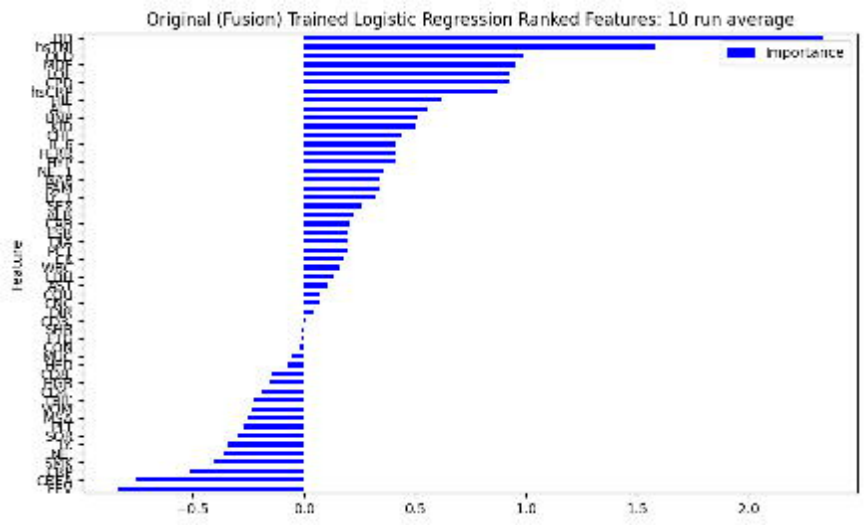

Preprocessing
Raw Data
Feature Data
Labels
Train-Test Split 80%-20%
Feature Training Data
Feature Testing Data
Training Labels
Testing Labels
Standard Scaling
Feature Training Data
Feature Testing Data
Training Labels
Testing Labels


Training and Predicting
Feature
Training
Data
Training
Labels
ML Model
Hyperparameter Search
and Model Training
Feature
Testing
Data
Tuned
Predictive
Model
Predicted
Testing Labels
0, 1, 0, 1…
True Testing
Labels
0, 1, 1, 0…


Omicron (Fusion) Trained Random Forest Ranked Features: 10 run average
feature
Importance
0.00
0.02
0.04
0.06
0.08
0.10


Original (Fusion) Trained Random Forest Ranked Features: 10 run average
Feature
Importance
0.00
0.02
0.04
0.06
0.08


Omicron (Fusion) Trained Logistic Regression Ranked Features: 10 run average
Importance
feature
−1
0
1
2
3
4


Original (Fusion) Trained Logistic Regression Ranked Features: 10 run average
Importance
feature
-0.5
0.0
0.5
1.0
1.5
2.0