**Peer review**

Tseng, Brian, Alejandro M Estrada, Nathan Hsu, Raaief Raaief, Honglun Xu, Tzu L Tseng, and Solayman H Emon. 2024. "Enhancing the COVID-19 Diagnostic for Patients Using Machine Learning and Voting Classifier Approaches." *Journal of High School Science* 8 (4): 331–49.

20% of 2000 is 400, yet your confusion matrices add up to 465 cases. Please explain.
Figure 2, a and b, needs consistency. For both figures make sure blue and orange refers to the same labels.

What is often neglected in such studies is that the data fed into the system (the PCR) in this case is also subject to sensitivity and specificity errors. For example, PCR COVID 19 data in a clinical setting is reported to be anywhere from 65% to 80% sensitive, i.e. it can detect 65 to 80 out of 100 as true positives
(see: https://doi.org/10.1371/journal.pone.0251661 and https://www.cap.org/member-resources/articles/how-good-are-covid-19-sars-cov-2-diagnostic-pcr-tests#:~:text=The analytic performance of PCR,specificity is near 100%25 also., this means that the model's prediction error is compounded due to the input sensitivity value. Please explain and discuss in the manuscript.

If your ensamble model is 80% sensitive, that means that the compounded sensitivity for the model (using 80% sensitive data for the PCR) is 63%. Assuming Coxs proportional hazards model. Therefore, your true sensitivity is 63%. Therefore, as a worst case scenario, you may not be isolating 37% of the people who may actually need to be isolated. If you plug these numbers, into the omni calculator, the # of weeks it takes for new cases to return to zero is 36 and 91 weeks for 1% infected versus 37% infected. Similarly, the # of weeks it takes for # infected people to return to zero is 90 verus 124 weeks for 1% infected versus 37% infected. This is a significant # of lives lost (assuming 1000 people die per day, 1000x 240 days is 240,000 additional people (quarter of a million) dead due to the 80% sensitivity of your model compounded with the 80% sensititivity of the PCR test in a clinical setting. Please discuss this in the manuscript. This is an important point because most authors (including yourselves) assume that the predictor variable as entered into the machine language models is 100% true - when, in actuality, it is not.

Going back to my earlier points, therefore, even if you were to obtain 99% recall and precision in your ML model, it would still only be as sensitive or specific as the PCR diagnostic test.

For all the models in which you used 5-fold cross validation, please present the accuracy, recall, precision and F1 values for each run in the manuscript.
There are 3 methods to mitigate class imbalance in reference 2 (RUS, ROS and SMOTE), which one did you use?
Please present a correlogram or a heat-map for all the features used and their Pearson correlation coefficients. This is standard for all AI ML manuscripts.

You mention 30 vital measurement in the text but table 1 presents only 18. What about the rest? Please present all 30 measurements. Present which of your models utilized which features? Did each model rank the feature in order of importance using one of the accepted methods such as recursive feature elimination, Permutation importance, LIME, SHAP ? You mention that your decision tree used GINI. What about the other models? Were all the

features used for the prediction or were only the top most explanatory features used? Please discuss and describe in the manuscript.

Did each model check for feature multicollinearity? If found, was a PCA used instead. Multicollinearity will not change prediction outcome but will affect the relative magnitude of the coefficients of the explanatory variables. This may provide incorrect information to decide which vital signs are important when formulating policy or mitigation procedures. Please describe and discuss in the manuscript.

---

## Detailed Response to Reviewer

Submission ID: 2767; Manuscript Title: "Enhancing the COVID-19 diagnostic for patients using machine learning and voting classifier approaches"

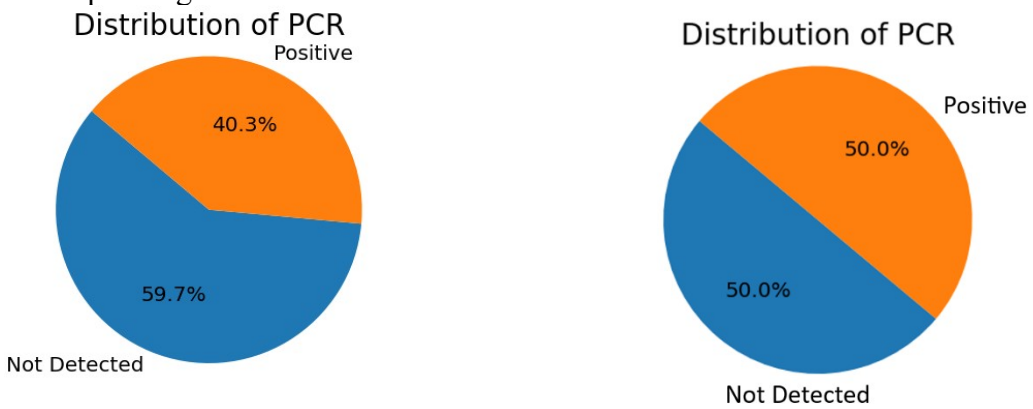
**Dear Reviewer:**

Thank you for allowing us to improve this manuscript. We deeply appreciate the reviewers' constructive comments and suggestions. Following these, we have diligently revised the previous manuscript, making improvements that have been highlighted in red in the revised version. Point-to-point responses to the comments are provided as follows.

### Reviewer #1

**Comment #1.1:** "Figure 2, a and b, needs consistency. For both figures make sure blue and orange refers to the same labels."

**Response:** Thank you for pointing out the issue. We apologize for the oversight. In the revised manuscript on updated Figure 2, we have ensured that the colors used are consistent with the corresponding labels.




(a) Imbalanced class distribution (b) Balanced class distribution (Oversampling)

Figure 2. Class distribution before and after data balancing

**Comment #1.2:** "Please present a correlogram or a heat-map for all the features used and their Pearson correlation coefficients. This is standard for all AI ML manuscripts."

**Response:** We sincerely appreciate the insightful suggestion. We agree that including a correlation plot can significantly enhance the clarity of identifying relations between variables. We extended our data analysis to add new correlation Figure 3 in the revised manuscript.

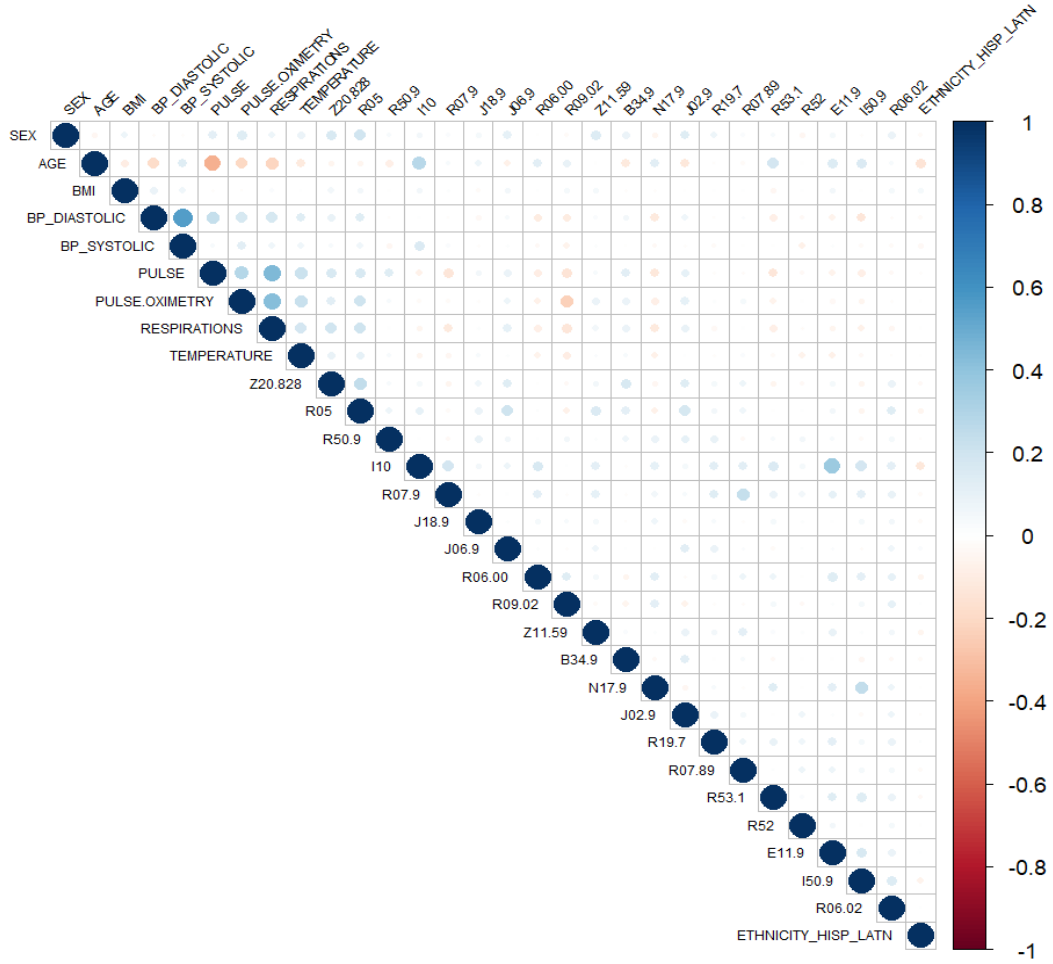



Figure 3. Correlation analysis among features

**Comment #1.3:** "You mention 30 vital measurements in the text but table 1 presents only 18. What about the rest? Please present all 30 measurements."

**Response:** Thanks for this feedback. In response to this feedback, we have updated the Table 1 including all 30 measurements with the description.

Table 1. Overview of the UTMB dataset

| Feature | Description |
|---|---|
| SEX (binary) | Gender of the patient |
| ETHNICITY (binary) | Self-reported ethnicity of the patient |
| AGE | Age of the patient in years |
| BMI | Body Mass Index, a measure of body fat based on height and weight |
| BP_DIASTOLIC | Diastolic blood pressure, measuring pressure in the arteries when the heart rests between beats |

| | |
|---|---|
| BP_SYSTOLIC | Systolic blood pressure, measuring pressure in the arteries when the heart beats |
| PULSE | Heart rate in beats per minute |
| PULSE.OXIMETRY | Oxygen saturation in the blood, as a percentage |
| RESPIRATIONS | Respiratory rate in breaths per minute |
| TEMPERATURE | A patient's temperature |
| Z20.828(binary) | Exposure to viral communicable eases |
| R05(binary) | Cough |
| R50.9(binary) | Unspecified fever |
| I10(binary) | Essential (primary) hypertension |
| R07.9(binary) | Unspecified chest pain |
| J18.9(binary) | Unspecified pneumonia |
| R06.9(binary) | unspecified abnormalities of breathing |
| R06.00(binary) | Dyspnea, unspecified |
| R09.02(binary) | Hypoxemia |
| Z11.59(binary) | Screening for other viral diseases |
| B34.9(binary) | Unspecified viral infection |
| N17.9(binary) | Acute kidney failure that is unspecified |
| J02.9(binary) | Acute pharyngitis, unspecified |
| R07.89(binary) | Other chest pain |
| R17.9(binary) | Hyperbilirubinemia without jaundice |
| R53.1(binary) | Weakness, diminished or absent energy and strength, or a lack of concentration |
| R52(binary) | Pain that is unspecified |
| E11.9(binary) | Type 2 diabetes mellitus |
| I50.9(binary) | Unspecified heart failure |
| R06.02(binary) | Shortness of breath |
| PCR (binary) | PCR test result for COVID-19, indicating virus detection status |

**Comment #1.4:** "Did each model check for feature multicollinearity? If found, was a PCA used instead. Multicollinearity will not change prediction outcome but will affect the relative magnitude of the coefficients of the explanatory variables. This may provide incorrect information to decide which vital signs are important when formulating policy or mitigation procedures. Please describe and discuss in the manuscript."

**Response:** We appreciate the reviewer's suggestion to analyze the feature multicollinearity for this manuscript. We acknowledged that presence of multicollinearity could affect the relative magnitude of the coefficients of the explanatory variables. To detect the multicollinearity among the features, we calculated the Variance Inflation Factor (VIF) value and present those with a new Figure 4 in the revised manuscript.

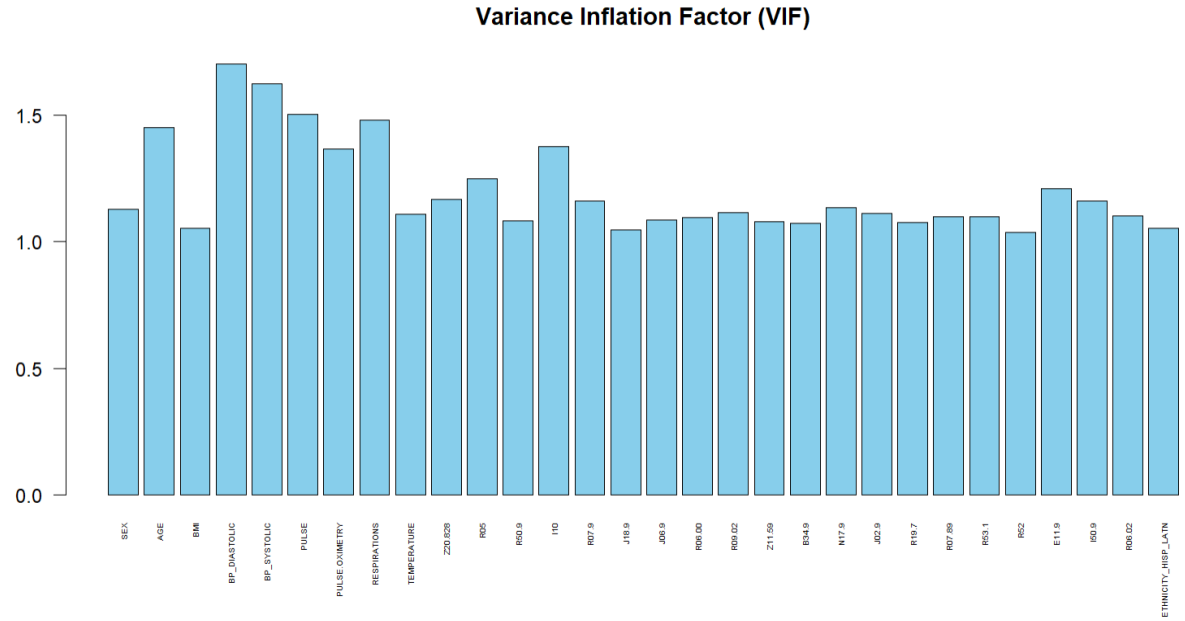



Figure 4. Assesement of multicollinearity among features

From Figure 4, we can observe that there is no critical multicollinearity (VIF < 2) found among the features. However, we compute the principal component (pc) for our further analysis as we have higher-dimensional situations (around 30 features).

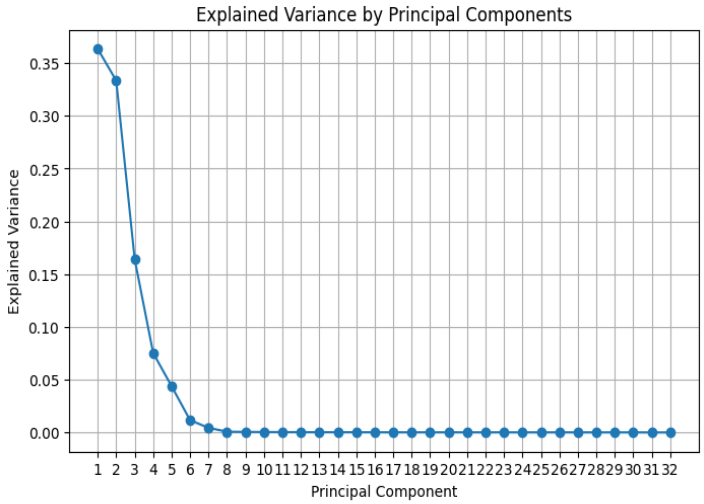



Figure 5. Analyis of principal components (PCs)

Although we didn't observe any problematic multicollinearity, with the concern of high dimensionality (~30 features) we further perform principal component analysis (Figure 5). We can observe that with seven principal components (pc=7), it would be possible to explain reasonable amounts of variance inside the data.

Additionally, in the result section, we add a new Table 6 for the comparison of performance metrics for the proposed voting classifier while taking into account both with and without utilization of the principal component analysis (PCA) technique.

**Table 6.** Comparison of voting classifier performance metrics with and without PCA technique

| Method | Dimension Reduction | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Voting Classifier | With PCA (n=7) | 0.745 | 0.731 | 0.724 | 0.738 |
| | Without PCA | 0.804 | 0.809 | 0.805 | 0.807 |

**Comment #1.5:** "How many epochs until convergence? or until loss was minimum for the models? Please present this data"

**Response:** Thanks for pointing out the valid concern. In the revised manuscript, we have added more description regarding this concern. The explanations could be found as follows:

- **SVM Convergence:** The convergence in SVM based on the optimization problem through finding maximum boundary/appropriate support vectors rather than iterating over epochs. When the solver algorithm converges on an optimal solution, the training process would be completed. The GridSearchCV performs cross-validation over different combinations of hyperparameters and selects the best set of parameters.

- **CART & RF Convergence:** Decision tree and Random Forests converge within the criterion named gini impurity rather than any epochs. The GridSearchCV iterates over the hyperparameter grid to find the best combination based on cross-validation.

- **ANN:** The training process of artificial neural network has a default setting for the number of epochs and loss function. The training process will continue until the model converges within maximum iteration number. As we are dealing with a classification problem, we utilized cross-entropy loss in the training scheme.

In the revised manuscript, we have added the description of this on 'Methodology' section.

**Comment #1.6:** "For all the models in which you used 5-fold cross validation, please present the accuracy, recall, precision and F1 values for each run in the manuscript."

**Response:** We appreciate this valuable suggestion. In response to this, we added a new Table 3 in the revised manuscript which will show the metrics results across cross-validation folds.

**Table 3.** Performance metrics of different models across cross-validation folds

| Model | Fold | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| SVM | fold1 | 0.715 | 0.731 | 0.710 | 0.723 |
| | fold2 | **0.723** | **0.722** | **0.737** | **0.730** |
| | fold3 | 0.713 | 0.715 | 0.713 | 0.713 |

| | | | | | |
|---|---|---|---|---|---|
| | fold4 | 0.710 | 0.711 | 0.710 | 0.709 |
| | fold5 | 0.713 | 0.715 | 0.713 | 0.713 |
| CART | fold1 | 0.678 | 0.660 | 0.662 | 0.673 |
| | fold2 | 0.680 | 0.691 | 0.679 | 0.701 |
| | fold3 | 0.712 | 0.671 | 0.663 | 0.682 |
| | fold4 | **0.690** | **0.707** | **0.665** | **0.686** |
| | fold5 | 0.658 | 0.680 | 0.667 | 0.688 |
| RF | fold1 | 0.742 | 0.742 | 0.742 | 0.742 |
| | fold2 | **0.763** | **0.756** | **0.788** | **0.772** |
| | fold3 | 0.749 | 0.752 | 0.749 | 0.748 |
| | fold4 | 0.720 | 0.721 | 0.720 | 0.720 |
| | fold5 | 0.699 | 0.700 | 0.699 | 0.698 |
| ANN | fold1 | 0.721 | 0.730 | 0.731 | 0.711 |
| | fold2 | 0.753 | 0.755 | 0.753 | 0.752 |
| | fold3 | **0.740** | **0.724** | **0.787** | **0.755** |
| | fold4 | 0.717 | 0.709 | 0.717 | 0.721 |
| | fold5 | 0.731 | 0.711 | 0.721 | 0.731 |

**Comment #1.7:** "There are 3 methods to mitigate class imbalance in reference 2 (RUS, ROS and SMOTE), which one did you use?"

**Response:** Thanks for pointing out the discrepancy. In the revised manuscript, we summarize the results obtained using three different data balancing methods such as Random Under-Sampling (RUS), Random Over-Sampling (ROS), and Synthetic Minority Over-sampling Technique (SMOTE) by the inclusion of a new Table 5.

**Table 5.** Performance metrics of voting classifier (VC) with different data balancing methods

| Method | Data Balance Method | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Voting Classifier | RUS | 0.762 | 0.770 | 0.803 | 0.784 |
| | ROS | 0.804 | 0.809 | 0.805 | 0.807 |
| | SMOTE | 0.781 | 0.756 | 0.720 | 0.749 |

**Comment #1.8:** "20% of 2000 is 400, yet your confusion matrices add up to 465 cases. Please explain."

**Response:** Thanks for investigating this critical point. We apologize for this inconsistency in the manuscript. Originally, the covid dataset has 1987 observations. In the Figure 1(a), we performed outlier analysis, then remove those outlier observations with IQR bounds. Therefore, we ended up with 1860 observations for model building. We split those observations with 75%-25% train-test split. In the manuscript we mistakenly mentioned 80%-20%. In the revised version, we fixed that inconsistency. Here are the details regarding the data distribution in training scheme.

| | |
|---|---|
| No. of Observations | 1860 |
| train-test split | 75%-25% |
| No. of train observations | 1395 |
| No. of test observations | 465 |

**Comment #1.9:** "You mention 30 vital measurements in the text but table 1 presents only 18. What about the rest? Please present all 30 measurements. Present which of your models utilized which features? Did each model rank the feature in order of importance using one of the accepted methods such as recursive feature elimination, Permutation importance, LIME, SHAP? You mention that your decision tree used GINI. What about the other models? Were all the features used for the prediction or were only the top most explanatory features used? Please discuss and describe in the manuscript."

**Response:** We appreciate for this important recommendation. In a previous response, we discussed about the principal component analysis (PCA) for the dimensionality reduction aspect. In the proposed model architecture, we utilized all the 30 vitals measurements. As we have access of a relatively smaller dataset (~2000), we decided to consider all the features for the model building purpose. However, we perform feature contribution analysis through Local Interpretable Model-agnostic Explanations (LIME) technique. In the revised manuscript, we added two new Figures (Figure 11 & 13) for this analysis purpose.

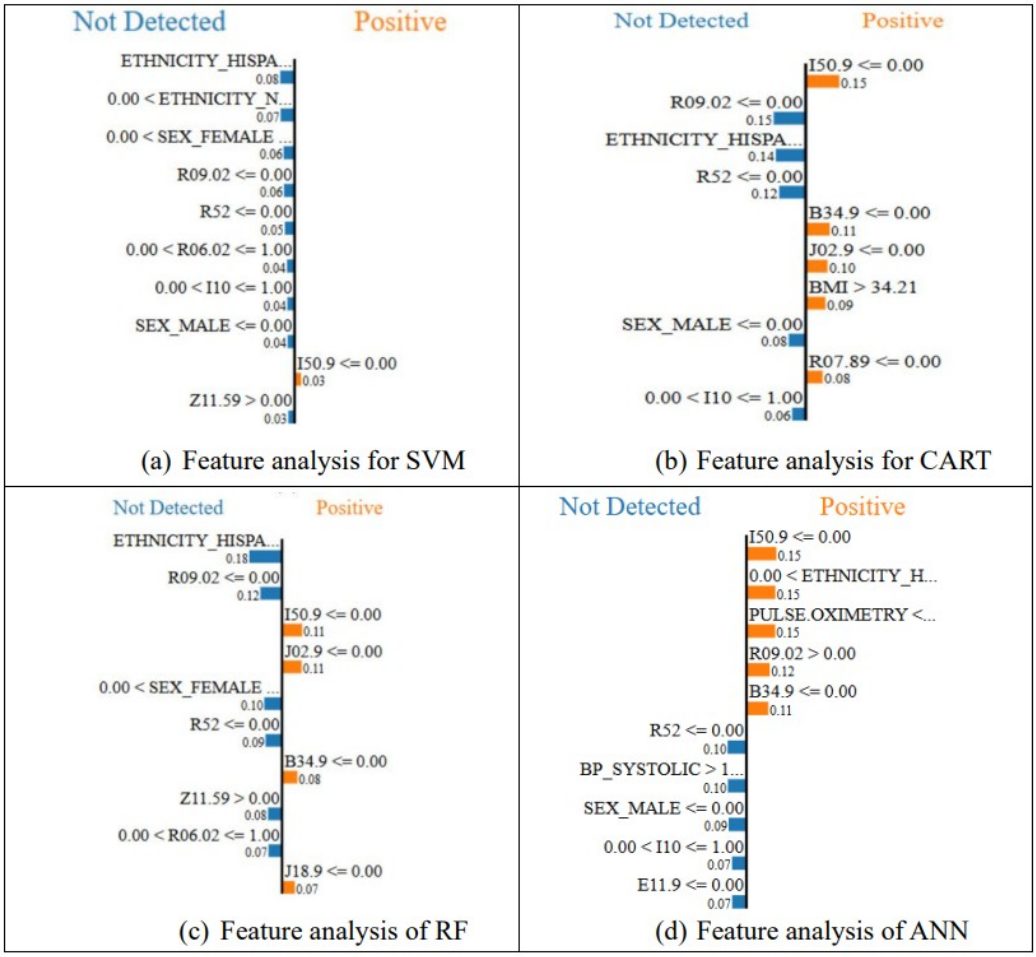



Figure 11. Feature contribution analysis for different machine learning models using LIME

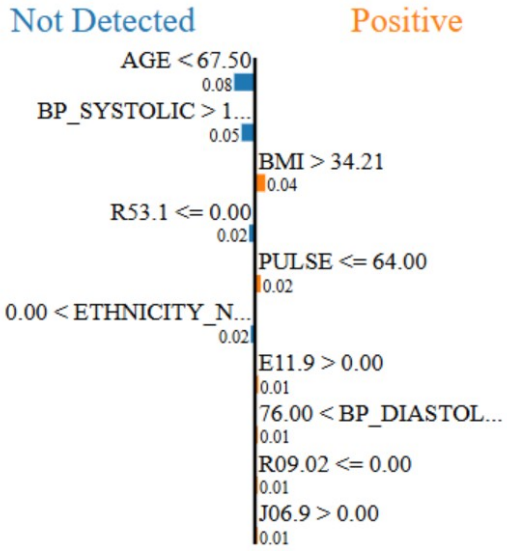



Figure 13. Feature contribution analysis for voting classifier (VC)

Regarding the question about the GINI Index, we utilized Gini Impurity as a criterion for all the tree-based methods we have applied in this use case. In the revised manuscript we have already highlighted this.

**Comment #1.10:** "Even if you were to obtain 99% recall and precision in your ML model, it would still only be as sensitive or specific as the PCR diagnostic test."

**Response:** Thanks for this thoughtful feedback. It would be a great learning opportunity through this critical analysis. We are fully agreeing that the inclusion of this critical analysis would add good value to the COVID-19 decision making process. In response to this suggestion, we mentioned about this critical point in the 'Discussion' section of the revised manuscript. Considered as a great learning material we added the two suggested articles about the compound sensitivity in the revised manuscript (Reference: 13 &14).

The updated 'Discussion' section in the revised manuscript could be found as:

**Discussion**

The application of machine learning models in COVID-19 diagnostics has shown considerable promise in enhancing both diagnostic accuracy and efficiency. This study employed five distinct supervised machine learning models: Support Vector Machine (SVM), Classification and Regression Tree (CART), Random Forest (RF), Artificial Neural Network (ANN), and a Voting Classifier. The results indicate that while individual models like SVM and ANN showed commendable performance, the Voting Classifier, which integrates the strengths of SVM, RF, and ANN, yielded the highest accuracy of 80.4% and an AUC of 85%. This ensemble approach capitalizes on the diverse strengths of multiple models, reducing the risk of overfitting and enhancing generalization. The Artificial Neural Network (ANN) achieved an accuracy of 73.9%, and an AUC of 0.72. However, the ANNs possibly could capture more complex pattern from data if the variation inside the dataset is larger enough. Although we found reasonable performance from the almost all the classification evaluation metrices, there is still have concern for model's erroneous predictions. model's

prediction error is compounded Due to the input sensitivity value, the machine learning model's prediction could be erroneous. In our COVID-19 analysis, we assumed that the data collection for 'PCR' don't have any sensitivity. However, In the real-life clinical setting, it is so obvious to present sensitivity among the collected data (13, 14). Therefore, if we consider the input data have sensitivity ($\partial$) and machine learning model's sensitivity ($\beta$), the actual sensitivity would be compounded with both. The compounded measurement would be $\partial \times \beta$, which would be different from whatever we compute in the conventional way. In the next iteration, we will emphasis on tracking the input data sensitivity during the data collection. Future work should focus on mitigating the compounded sensitivity issue, expanding the dataset, and incorporating additional features to further enhance model performance. Additionally, exploring other ensemble techniques and advanced machine learning algorithms could provide even greater improvements in diagnostic accuracy and reliability. The continued evolution and refinement of these models hold promise for revolutionizing COVID-19 diagnostics and potentially other areas of medical diagnostics, ultimately contributing to better healthcare outcomes.

---

We would like to sincerely thank the thoughtful and constructive feedback. Your valuable suggestions have greatly helped us enhance the clarity and depth of this work. We remain grateful for your continued guidance and are happy to address any additional comments.