An Explainable AI Approach for Diagnosis of COVID-19 using MALDI-ToF Mass Spectrometry

Venkata Devesh Reddy Seethi\textsuperscript{a}, Zane LaCasse\textsuperscript{a}, Prajkta Chivte\textsuperscript{a}, Joshua Bland\textsuperscript{b}, Shrihari S. Kadkol\textsuperscript{b}, Elizabeth R. Gaillard\textsuperscript{b}, Pratool Bharti\textsuperscript{c} and Hamed Alhoori\textsuperscript{d}

\textsuperscript{a}Northern Illinois University, 1425 W Lincoln Hwy, Dekalb, Illinois, 60115, USA  
\textsuperscript{b}University of Illinois Chicago, 840 S Wood St, Chicago, Illinois, 60612, USA  
\textsuperscript{c}Intel Corporation, 2501 NE Century Blvd, Hillsboro, Oregon, 97124, USA

\textbf{ARTICLE INFO}

Keywords:  
COVID-19 testing  
Explainable AI  
Machine learning  
RT-PCR test  
Mass spectrometry  
MALDI-ToF

\textbf{ABSTRACT}

The severe acute respiratory syndrome coronavirus type-2 (SARS-CoV-2) caused a global pandemic and immensely affected the global economy. Accurate, cost-effective, and quick tests have proven substantial in identifying infected people and mitigating the spread. Recently, multiple alternative platforms for testing coronavirus disease 2019 (COVID-19) have been published that show high agreement with current gold standard real-time polymerase chain reaction (RT-PCR) results. These new methods do away with nasopharyngeal (NP) swabs, eliminate the need for complicated reagents, and reduce the burden on RT-PCR test reagent supply. In the present work, we have designed an artificial intelligence-based (AI) testing method to provide confidence in the results. Current AI applications for COVID-19 studies often lack a biological foundation in the decision-making process, and our AI approach is one of the earliest to leverage explainable AI (X-AI) algorithms for COVID-19 diagnosis using mass spectrometry. Here, we have employed X-AI to explain the decision-making process on a local (per-sample) and global (all samples) basis underscored by biologically relevant features. We evaluated our technique with data extracted from human gargle samples and achieved a testing accuracy of 94.12%. Such techniques would strengthen the relationship between AI and clinical diagnostics by providing biomedical researchers and healthcare workers with trustworthy and, most importantly, explainable test results.

\textbf{1. INTRODUCTION}

The virulent, fast-spreading, and mutating nature of SARS-CoV-2 caused a worldwide pandemic with more than 765 million people infected, and 6.9 million lives were taken in the span of 3 years since emerging (WHO, 2023). This pandemic has challenged the ability of healthcare systems to triage patients and overwhelmed capacities. Beyond healthcare, societal impacts have been apparent, and major sectors of the global economy have been affected (Sohrabi et al., 2020). Given the severity of the pandemic and the emergence of new variants of SARS-CoV-2 (NCIRD, 2023), curbing the infection caused by the virus has become a major concern for all nations. Several protocols have been implemented, such as vaccinations, mask mandates, and lockdowns, to reduce the spread of the virus. Frequent and rapid surveillance testing, nonetheless, remains the key critical factor in reducing the transmission of the virus (Nicola et al., 2020). As COVID-19 exhibits a broad range of clinical manifestations, from mild fever to life-threatening conditions, it is critical to implement testing protocols for its early diagnosis in order to curb viral transmission. Molecular detection of SARS-CoV-2 predominantly relies on the current gold standard Real-Time Polymerase Chain Reaction (RT-PCR) technique (Filchakova et al., 2022) for viral RNA identification. Despite ongoing COVID-19 testing advancements, there is still a significant trade-off between testing specificity and complexity. Although RT-PCR protocols have achieved high specificities of 95.2% (Habibzadeh et al., 2021), complete reliance on RT-PCR exerts a strain on laboratory reagents for the testing process (Carter et al., 2020) and the availability of skilled personnel to conduct the relatively labor-intensive process. Various other alternatives for detecting SARS-CoV-2 nucleic acids, such as isothermal amplification assays, amplicon-based metagenomic sequencing, hybridization using microarray and CRISPR-based technologies, are simultaneously being developed to meet the unprecedented demands for a rapid but reliable testing platform (Carter et al., 2020; Feng et al., 2020; Habibzadeh et al., 2021; Lasserre et al., 2022). Approximately 400 rapid diagnostic tests are commercially available, amongst which lateral flow devices are the most prominent ones and provide up to 70% sensitivity to detect antibodies (Valera et al., 2021).

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-ToF) has become one of the leading proteomic tools due to its minimal sample preparation, tolerance for matrix contaminants, and rapid analysis providing multidimensional information (Preiàno et al., 2021; Sivanesan et al., 2022; Spick et al., 2022). Previous research reported a MALDI-ToF-based method for profiling proteins in gargle samples where both host and viral proteins are observed. This study included samples from people whose COVID-19 status was verified with RT-PCR testing. Utilizing straightforward area under the curve (AUC) analysis of protein signals in the MALDI-ToF spectra, the method achieved overall...
above 90% specificity and sensitivity (recall) rates (Chivte et al., 2021).

In this study, we utilized mass spectra data for gargle samples obtained through MALDI-ToF process from Chivte et al. (2021), along with newly collected samples. An in-depth look at the spectra, as illustrated in Figure 1, shows that they are fairly complicated, with a multitude of signals detected from both host immune response proteins and viral proteins. The amounts of these proteins change over the course of an infection’s lifecycle, further confounding the data in the spectra. Thus, we now report on an alternative analysis using AI and X-AI algorithms.

The complexity of patterns learned by AI systems often results in opaque and uninterpretable decision-making processes, creating black-box systems. In clinical settings, black-box systems pose an impossible task for domain experts and health practitioners to understand the factors contributing to a test’s outcome. Therefore, we tackle this problem using X-AI techniques that document the decision-making process in transferring a mass spectral dataset into a COVID-19 diagnosis. Although the AI system’s decision-making process is complex, the explanations are easy to understand, assisting clinicians in interpreting the rationale behind the diagnosis. This characteristic is analogous to a real-time diagnosis, where doctors explain to patients what markers lead to their final diagnosis. Therefore, ensuring model interpretability is critical, as it fosters transparency and instills trust in our AI system.

2. LITERATURE REVIEW

Machine learning algorithms have become an integral part of research in COVID-19 diagnostic protocols (Khan et al., 2021). As discussed in Section 1, numerous protocols have investigated different sample types, feature engineering techniques, and machine learning algorithms. In previous mass spectrometry studies that use NP swabs, Nachigall et al. (2020) collected 362 samples that were analyzed by MALDI-ToF Mass Spectrometry (MS) in the mass-to-charge range $3 - 15.5$ kDa. Spectral peaks were extracted by filtering local maxima, followed by correlation-based and impurity-based feature selection methods to identify 88 important peaks that discriminate between COVID-19 positive and negative samples. They achieved 93.9% accuracy with the support vector machine (SVM) algorithms (Noble, 2006). Rocca et al. (2020) collected 311 samples and conducted several statistical tests to select characteristic peaks, and their k-nearest neighbor classifier achieved an accuracy of approximately 90%. In other work, Tran et al. (2021) collected 199 samples, extracted peaks, and optimized their AI solution using a machine intelligence learning optimizer (MILO) (Rashidi et al., 2022). MILO builds an end-to-end pipeline that automatically selects the best features and tunes hyperparameters for algorithms (Rashidi et al., 2022). The results of MILO showed that extreme gradient boosting (XgBoost) (Chen & Guestrin, 2016), and the deep neural network (DNN) (LeCun et al., 2015) performed best with 96.6% and 98.3% accuracies, respectively. In another study, Deulofeu et al. (2021) collected two datasets with 148 and 66 samples using different sample processing techniques. The authors generated 5-50 principal components from the peaks.
dataset using principal component analysis (Wold et al., 1987) and reported the accuracies as 70.3% with XGBoost and 88.4% with SVM for their two datasets, respectively.

Alternative to NP swabs, Yan et al. (2021) collected 298 serum samples. In feature processing, they selected 50 peaks where 20, 20, 10 peaks were determined using LASSO (Tibshirani, 1996), partial least-squares-discriminant analysis (Ståhle & Wold, 1987), and recursive feature elimination with cross-validation (Guyon et al., 2002), respectively. As a next step, they conducted empirical analysis, picked 25 relevant peaks from the 50 peaks, and found that logistic regression gave the best accuracy of 99%. Costa et al. (2022) explored the efficacy of saliva samples in COVID-19 testing and collected 360 samples. They further extracted and selected nine statistically significant peaks with the Wilcoxon rank sum test and Benjamini–Hochberg correction (Benjamini & Hochberg, 1995). Lastly, their SVM algorithm exhibited the best performance with accuracy ranging from 64.3% – 88.9%.

While several studies have used NP-swabs, blood or blood-based specimen type coupled with X-AI to diagnose COVID-19, our study focused on saliva sampling as it is non-invasive, convenient, comfortable across all age groups, and provides an option of self-sampling. Moreover, blood is a more complex biological fluid than saliva, which increases the cost of its storage and time for processing during the detection of biomarkers (Abraham et al., 2012; Esser et al., 2008; Lasisi & Lawal, 2019).

Researchers have also adopted AI approaches driven by X-AI analysis (Salahuddin et al., 2022) in COVID-19 diagnosis for different datasets such as blood samples (Alves et al., 2021; Gong et al., 2022; Rostami & Oussalah, 2022; Thimoteo et al., 2022; Yan et al., 2021), CT scans (Basu et al., 2022; Pennisi et al., 2021; Ullah et al., 2022; Ye et al., 2021), and X-Rays (Ardakani et al., 2020; Hu et al., 2022; Li et al., 2022; Ozturk et al., 2020; Sharma et al., 2022; Vinod et al., 2021; Wen et al., 2022). Rahman et al. (2021) used multimodal data from X-Rays, CT scans, cough audio data, and facial recognition to create an ensemble pipeline with multiple point-of-care premise tools to diagnose COVID-19. Some studies also explored forecasting COVID-19 cases by learning the time-series relationship between historical data, such as transmission and recovery rates, to forecast case numbers or chances of getting infected with COVID-19 (Leung et al., 2021; Nadimi-Shahraki et al., 2022; Nguyen et al., 2022). Smith & Alvarez (2021) identified mortality factors of patients in a hospital by training a model on tabular data with features such as age, gender, and lymphocyte count and presented the features that had a high contribution towards a patient’s mortality using X-AI techniques. Furthermore, explainable AI has been used in spectrometry datasets for biomarker discovery (Tideman et al., 2021) and cell classification (Xie et al., 2020) and has also shown potential in learning age group-dependent features in COVID-19 (Liangou et al., 2021).

In summary, previous X-AI studies are based on datasets that are tabular or 2D images. Tabular data generally consist of rows and columns, while image data can be either a 2D X-ray or CT scan. By looking at previous studies in X-AI, it is apparent that SHAP (Lundberg et al., 2020, 2018) is the best-known method for tabular datasets and decision trees (Safavian & Landgrebe, 1991) for building transparent models with tree-based explanations (Alves et al., 2021). On the other hand, for image datasets, GRAD-CAM (Selvaraju et al., 2017) and SCORE-CAM (Wang et al., 2020) techniques are more popular. In this study, we demonstrate an X-AI framework that converts mass spectrometry data into tabular data, trains an AI model, and generates explanations using multiple X-AI algorithms, which have been proponents in making AI models study trustworthy.
We have compiled several studies that incorporated MALDI-ToF analysis in Table 3 and compared the instrument, sample type, data type, sample processing time, dataset size, feature engineering technique, X-AI model, AI model, and accuracy. It is important to note that each study employs data that have different participants and different protocols. We aim to understand the machine-learning approaches used for different sample processing techniques. Other sample-related details as discussed by authors have not been discussed here (Sivanesan et al., 2022; Spick et al., 2022).

3. CONTRIBUTIONS

In our current work, we incorporate one of the newest sample collection processes known for cost-effectiveness and ease of implementation. We also present a thorough evaluation of our feature processing technique as well as from the previous mass spectrometry literature. Furthermore, we demonstrate that the diagnosis becomes feasible and accurate with the integration of AI algorithms. We additionally interpret the AI algorithm to provide informed explanations of diagnosis, leverage a cost-effective and easy-to-use MALDI-ToF MS approach, and use a non-invasive specimen type (saliva). We designed our AI approach to learn the relationships between the intensity of the protein signals and the outcome of an RT-PCR test. Our experimental setup, as illustrated in Figure 2, encompasses multiple stages: data collection, preprocessing, feature engineering, ML model training, and model interpretation. The first step in our experimental setup involves acquiring gargle samples, which are subsequently analyzed using MALDI-ToF to generate mass spectra. We then transform data by calculating the area under the curve (AUC) and employ the ratio of AUCs for feature engineering. Next, we apply anomaly detection techniques to filter the dataset to identify and filter any outliers before training the machine learning model. In the final stages, we evaluate feature importance, global and local model interpretation, and finally present an explainable AI framework that encapsulates our systematic workflow to generate the outcome along with explanations for the outcome on unseen samples.

### Table 1

| Biomarker | Mass Range (m/z) | Potential Identity |
|-----------|------------------|--------------------|
| B₀        | 27900 – 29400    | Human Protein-1 (HP1) |
| B₁        | 55500 – 59000    | Human Protein-2 (HP2) |
| B₂        | 66400 – 68100    | Viral Protein-1 (VP1) |
| B₃        | 78600 – 80500    | Viral Protein-2 (VP2) |
| B₄        | 111500 – 115500  | Human Protein-3 (HP3) |

Potential biomarker ranges and their descriptions.

### Table 2

| Feature | Description | Potential Identity |
|---------|-------------|--------------------|
| R₀      | AUCA_P/HAPA_C | HP1/HP2 |
| R₁      | AUCA_P/HAPA_C | HP1/VP1 |
| R₂      | AUCA_P/HAPA_C | HP1/VP2 |
| R₃      | AUCA_P/HAPA_C | HP1/HP3 |
| R₄      | AUCA_P/HAPA_C | HP2/VP1 |
| R₅      | AUCA_P/HAPA_C | HP2/VP2 |
| R₆      | AUCA_P/HAPA_C | HP2/HP3 |
| R₇      | AUCA_P/HAPA_C | VP1/VP2 |
| R₈      | AUCA_P/HAPA_C | VP1/HP3 |
| R₉      | AUCA_P/HAPA_C | VP2/HP3 |

Ratio features generated from AUC values of five potential biomarkers.

4. DATASET PREPARATION

4.1. Data Collection

We collected the samples from the students at Northern Illinois University in August and September 2020. Ethical approval was issued by the Institutional Review Board. All participants provided written consent before participating in the research study. Students who consented to participate in the research study were asked to provide a water gargle sample at the same time as the NP swab sample. After sample collection, we processed the samples as described below while NP swabs were subjected to RT-PCR testing conducted near the University. For each gargle sample, we used the outcome of the RT-PCR test with an NP swab as ground truth in our analyses. We analyzed 152 gargle samples, including 60 COVID-19 positives and 92 COVID-19 negatives.

In brief, participants gargled 10 mL of bottled spring water for 30 seconds and deposited the gargled sample in a 50 mL conical centrifuge tube. We then filtered the gargled sample with a 0.45µm syringe filter. Following this, we used acetone precipitation to concentrate proteins, centrifuged the sample for 30 minutes, and resuspended the pellet in a reconstitution buffer for a further 15 minutes. After incubation, we spotted each sample on a target plate, air-dried for 20 minutes, and transferred the plate to a Shimadzu AXIMA Performance mass spectrometer (Shimadzu, 2022), which was calibrated daily with a protein standard. On a single MALDI-ToF target plate, 394 samples can be spotted and it takes 4 minutes per sample to run on the instrument. Ions were detected in a positive-ion, linear detection mode over a range of 2000 – 200000 m/z. Lastly, we corrected the baseline of all the generated spectra using the Shimadzu Launchpad software obtained from the manufacturer (Shimadzu, 2022) and normalized the ion count intensities with the parent peak of the protein standard calibrant. Overall, our process from sample collection to MALDI-ToF analysis, takes less than 1.5 hours.
| Study                  | Instrument   | Data Type      | Sample Type | Sample Processing Time | Dataset Size | Feature Engineering | XAI Model     | AI Model   | Accuracy       |
|------------------------|--------------|----------------|-------------|------------------------|--------------|---------------------|---------------|------------|----------------|
| Nachtigall et al. (2020) | MALDI-ToF    | Mass spectra   | Nasal swabs | —                      | 362          | Selected peaks      | SVM           | —          | 93.9%          |
| Rocca et al. (2020)    | MALDI-ToF    | Mass spectra   | Nasal swabs | —                      | 311          | Selected peaks      | KNN           | —          | 90%            |
| Deulofeu et al. (2021) | MALDI-ToF    | Mass spectra   | Nasal swabs | —                      | 214          | PCA on peaks        | XgBoost SVM   | —          | 70.3% - 88.4% |
| Tran et al. (2021)     | MALDI-ToF    | Mass spectra   | Nasal swabs | < 1 hour               | 199          | Selected peaks      | XgBoost and DNN | —          | 98.3%          |
| Garza et al. (2021)    | ESI-MS (MassSpec Pen) | Mass spectra   | Nasal swab  | —                      | 171          | Selected peaks      | —             | —          | 78.4%          |
| Yan et al. (2021)      | MALDI-ToF    | Mass spectra   | Blood serum | —                      | 298          | Selected peaks      | LR            | —          | 99%            |
| Chivte et al. (2021)   | MALDI-ToF    | Mass spectra   | Gargle      | 1.5 hr                 | 60           | AUC                 | —             | —          | 91.5% - 96.7% |
| de Almeida et al. (2022)| MALDI FT-ICR| Mass spectra   | Saliva      | > 5.5 hours            | 149          | Selected peaks      | SVM           | —          | 86.7% - 95.6% |
| Thimoteo et al. (2022) | —            | Clinical variables | Blood test  | —                      | 1500         | Physiological measures | SHAP        | RF         | 70% - 90%      |
| Lazari et al. (2022)   | MALDI-ToF    | Mass spectra   | Saliva      | —                      | 232          | Filtered peaks      | —             | SVM        | 67.8% - 88.5% |
| Ullah et al. (2022)    | —            | X-Ray CT scan  | Chest scan  | —                      | 1646, 2481   | Contrast limited adaptive histogram equalization | GRAD-CAM | ensemble | 98.5%          |
| Xu et al. (2023)       | —            | CT scan        | Chest scan  | —                      | 194922       | Bicubic interpolation | —           | DenseNet121 with similarity regularization | 99.2%        |

Table 3
Literature survey exploring various platforms coupled with AI for the diagnosis of COVID-19.
4.2. Potential Biomarker Selection

The presence or absence of intensities in the collected mass spectra can be correlated to the presence or abundance of proteins within each specimen. The spectra collected represent a multitude of proteins (in this case, after acetone precipitation) that are in the range of 2,000 – 200,000 m/z. With MALDI-ToF MS, we measure large molecules that have a distribution over a range of m/z values instead of a single m/z value. While we aimed to distinguish between a positive and negative sample, we selected m/z ranges identified in Chivte et al. (2021) that were observed to change given the COVID-19 status by RT-PCR.

An overlay of mass spectra highlighting these differences for a COVID-19 positive and negative sample is shown in Figure 1. Note that Figure 1 only shows a segment of the spectra where potential COVID-19 biomarkers are present. These biomarker peaks were selected to potentially include both human proteins (i.e., immunoglobulin heavy chain, amylase, asymptomatic biomarker) and SARS-CoV-2 viral proteins (Spike protein subunits S2 and S1). For the sake of simplicity, here, these ranges were named Human Protein-1 (HP1), Human Protein-2 (HP2), Human Protein-3 (HP3), and Viral Protein-1 (VP1), Viral Protein-2 (VP2), respectively. We provide a summary of the biomarkers and their mass ranges in Table 1.

4.3. Feature Engineering

The performance of supervised machine learning (ML) algorithms depends on the quality and the dimensionality of the features used in training. In general, an AI model becomes more robust as the number of features increases and weakens due to overfitting (McLachlan, 2004). We thus conduct a holistic analysis of different feature processing techniques such as raw spectral data, peaks in spectra, statistical features, the area under the curve (AUC), the ratio of AUC features, and statistical hand-crafted features. We evaluated only one feature processing technique at a time without combining several feature processing techniques for two reasons: to keep the dimensionality of data as small as possible and to make the feature interpretation easy to understand. We implemented all feature engineering techniques in Python 3.8 using libraries such as NumPy (Harris et al., 2020), Pandas (McKinney, 2011), and SciPy (Virtanen et al., 2020). Below, we delineate all the feature engineering techniques experimented with in our study.

4.3.1. Raw spectral data

We had directly used the raw readings of all five chosen m/z ranges within the spectra, which resulted in 12600 features (the combined bandwidth of all five biomarker ranges) for each sample. However, this could potentially overfit the ML models. We have also attempted binning the spectra using window sizes such as 50 and 100; however, we did not see any improvement in performance using binning. Therefore, we extracted features that balanced being easy to compute but also had enough information to discriminate between the positive and negative samples.

4.3.2. Statistical features

We calculated simple statistical features for each biomarker range, such as minimum, maximum, standard deviation, variance, skewness, kurtosis, and the number of peaks. Therefore, since we have five biomarkers and seven features, we generated a total of $5 \times 7 = 35$ features. The statistical features extract the variables from the raw data and encapsulate the data in fewer dimensions, which reduces the burden on ML algorithms to learn patterns from the data.

4.3.3. Area Under the Curve (AUC)

As discussed in Section 4.2, the presence of a viral or human protein can be measured by the intensity of the spectral peak over a narrow distribution of m/z values and cannot be localized to a single point. Hence, AUC (Tallarida & Murray, 1987) is more suitable over local maxima for representing an effective viral or human protein signal as it aggregates the intensities of all points in the range. Moreover, Chivte et al. (2021) established AUC as a reliable feature for distinguishing COVID-19 positive and negative samples using gargle samples analyzed via mass spectrometry. We use the biomarkers from Chivte et al. and describe them as $B_0$–$B_4$, as shown in Table 1.

4.3.4. Ratio of the AUC values

Although the AUC captures the intensity of a biomarker, we observed a wide variance from one sample to another. This happens mainly due to two reasons. First, each sample has potentially different amounts of human/viral protein present. Second, information regarding exposure to the virus or the time course of infection was impossible to collect. Such differences in immune response and viral load would be reflected in the biomarker intensities. We tested different arithmetic combinations such as subtraction, addition, ratio, and product for each sample with all possible combinations of different biomarker pairs. Regardless, we observed that the ratio of AUC between pairs of biomarkers showed a consistent pattern for positive and negative samples (shown in Table 2).

4.4. Data Preparation

Initially we used stratified five fold cross validation to evaluate different machine learning algorithms for the various feature processing techniques listed above. For our final results, however, we used train-test split as we aim interpret the model using XAI techniques. partitioned 152 samples from our MALDI-ToF dataset using a train-test-split strategy with a 70 : 30 train-test-split ratio to form two disjoint and independent splits. This minimizes data leakage from classifier training into calibration and performance assessment. Within the train and test splits, we kept the ratio of the count of positive to negative samples constant to maintain class balance. Our classifier uses the train-split dataset using features in Table 2 (containing 41 COVID-19 positive and 65 COVID-19 negative) and learns the COVID-19 prediction task. Then, we further calibrated the classifier using the same train-split dataset. As a final step, we assessed the performance of the trained classifier on the unseen test-split
dataset (containing 19 COVID-19 positive and 27 COVID-19 negative samples).

5. METHODS

In our study, we utilized the RT-PCR outcome as the ground truth for our dataset. However, our dataset is prone to outliers for several reasons. First, the RT-PCR test results might not always be accurate. Another factor is the potential of some added noise during the sample processing stage, affecting the dataset’s consistency and model performance due to outliers. To address this issue, we employed the Isolation Forest (IF) (Liu et al., 2008, 2012) technique to identify potential outliers. IF is an unsupervised learning algorithm that efficiently detects outliers by isolating data points in a random decision tree structure. It assigns an anomaly score to each data point based on the average path length from the root node to the terminating node, effectively distinguishing outliers from normal observations. We identified and removed 20 outliers in our dataset using IF and then used stratified 5-fold stratified cross-validation on various machine learning algorithms and feature engineering techniques as shown in Figure 3. Removing outliers from our dataset helped enhance the reliability of our models and ensured a more robust analysis.

We chose the random forest (RF) classifier (Breiman, 2001) after assessing several machine learning algorithms such as gaussian naive bayes (GNB) (Rish, 2001), logistic regression (LR) (Hosmer et al., 2013), k-nearest neighbors (KNN) (Muñoz et al., 2009), decision tree (DT) (Safavian & Landgrebe, 1991), extreme gradient boosting (XGB) (Chen & Guestrin, 2016) and deep neural networks (DNN) (LeCun et al., 2015). Figure 3 shows the accuracy of the machine learning models when tested with different feature engineering techniques. We observed that neural networks (NN) had good performance when trained on raw spectral data. However, we chose RF over NN as there is potential to overfit while using raw data of high dimensions for training on small dataset. Moreover, RF is easier to interpret, requires less training time, and is widely used in medicine-related applications (Alves et al., 2021; Shiri et al., 2022; Vinod et al., 2021).

The objective of RF or any other ML algorithm is to train a classifier function \( f \) that learns the relationships between the ratio features generated from training samples \( x_{train} \) and the outcome of the associated COVID-19 RT-PCR test \( y_{train} \). After training, we calibrated the RF and evaluated on new/unseen saliva samples \( x_{test} \) to evaluate the precision, recall, and F1-score metrics of the trained classifier’s prediction \( f(x_{test}) = y_{test} \) as compared to the corresponding RT-PCR result. We then leveraged X-AI algorithms to illustrate explanations for the outcomes of the RF algorithm on both a global (all-samples) and local (per-sample) basis. As a final step, we consolidated the outcomes of the RF model and explanations of X-AI algorithms into an easy-to-interpret 3-stage framework for COVID-19 diagnosis.

We used several Python libraries for implementing AI and X-AI algorithms. For instance, we used Scikit-Learn (Pedregosa et al., 2011) for all machine learning algorithms and Tensorflow (Abadi et al., 2016) for neural networks. For X-AI algorithms, we used Scikit-learn and SHAP library (Lundberg et al., 2020; Lundberg & Lee, 2017).

5.1. Random Forest (RF) Classifier

RF is an ensemble classification algorithm that combines several tree predictors to build a forest classifier. Each tree in the forest trains from a randomly sampled vector from the input features. These random vectors are sampled independently from the same distribution in the predictor space (Breiman, 2001). The outcome of a trained forest is the majority vote taken in as the mean or median of the predictions from all the independently trained trees. Additionally, in the RF algorithm, we can fine-tune the parameters such as the number of trees (also known as estimators) in the forest, methods of bootstrapping, and the properties of trees.

Figure 3: Five fold stratified cross-validation accuracy scores for machine learning models (shown on the x-axis) and feature engineering techniques (shown on the y-axis).
Table 4
Hyperparameters of our RF classifier.

| Hyperparameter       | Value   |
|----------------------|---------|
| Number of estimators | 100     |
| Bootstrap            | Enabled |
| Splitting criterion  | Gini impurity |
| Maximum depth of tree| 3       |
| Minimum samples per split | 2     |
| Minimum samples per leaf | 1     |

If bootstrapping is enabled, random samples are drawn from the training set with replacements to train decision trees and take the majority vote. Furthermore, we can tune the core properties of each decision tree, such as maximum depth, the maximum number of leaves, splitting criterion, minimum samples per split, and minimum samples per leaf, to optimize the performance.

Our first step in training the RF classifier was to find optimized hyperparameters by leveraging grid search with stratified 5-fold cross-validation on the training dataset. Grid search picks hyperparameter configurations from our defined grid of hyperparameters, wherein we cataloged each configuration with its cross-validation accuracy score. Finally, the hyperparameter configuration with the best score (shown in Table 4) was selected to train the RF algorithm on complete training data, which is further evaluated and interpreted with the test-split data.

5.2. Random Forest Calibration

Let our trained RF model \( f \) compute the probability \( P \) for a sample \( x_i \) to test positive, then \( P(f(x_i) == COVID-19\ positive) \) is computed as the fraction of trees in the forest that vote for COVID-19 positive. The final outcome is then predicted as COVID-19 positive if \( P(f(x_i)) > 0.5 \). While the final outcome helps in making a binary decision, the probability can help medical practitioners to understand the certainty of having the COVID-19 disease. Furthermore, the binary outcome of COVID-19 positive or COVID-19 negative tests does not help diagnose early infection and recovery stages. Rather than a simple “yes” or “no” answer, knowing the probability of having COVID-19 is crucial as it helps to understand the intermediary stages in the disease cycle.

Although RF achieves good performance, the predicted probabilities are sometimes out of step with their actual probabilities (i.e., the ratio of positive samples and the entire dataset). A classifier indicates ideal calibration if the predicted and actual probabilities are synchronized for the entire probability range \( 0 - 1 \) (shown as a gray dotted line in Figure 4). For example, if an ideally calibrated classifier predicts 0.8 probability for COVID-19 positive, then eight of ten samples should be COVID-19 positive (meaning actual probability is \( 8/10 = 0.8 \)). However, in a real-life scenario, the classifiers require additional calibration to refine their probability scores.

Table 5
Samples in each region segregated with thresholds on probability after removing outliers (for calibrated classifier).

![Calibration curves for un-calibrated and calibrated RF classifier. The gray dotted line represents an ideally calibrated classifier.](image)

The most common calibration methods include isotonic regression (Menon et al., 2012) and Platt scaling (Platt, 1999). Isotonic regression fits a step-wise regression function to the predicted probabilities and maps them to the actual probabilities, which usually works well for larger datasets. On the other hand, Platt scaling is compatible with even smaller datasets such as ours. Therefore, we used Platt scaling that adds a one-dimensional logistic regression (Hosmer et al., 2013) function to the output of RF that calibrates RF’s predicted probabilities with a sigmoid function (Goodfellow et al., 2016).
After calibrating RF, we investigated the calibration quality by plotting a calibration curve (as shown in Figure 4) using the train-split dataset. The calibration curve has predicted probabilities (on the x-axis) and actual probabilities (on the y-axis). Note that the probabilities were binned into five bins to make the curve less noisy and more readable. We observed that the probabilities were well-calibrated for calibrated RF compared to the original RF. We also graphed the predicted probabilities for each sample in the train-split and the test-split datasets in Figure 5 to get a more in-depth view of the probability scores for each sample. In this plot, all the positive samples are colored red and negative samples with blue. All samples that were identified as anomalies/outlier by IF are marked with a grey triangle on top. According to the train-split dataset’s probability scores, we found that setting a gatekeeper threshold for probability resulted in better performance. Hence, based on the train-split dataset, we set the threshold to predict COVID-19 positive when probability (P) is \( P \geq 0.7 \) and \( 0.25 \geq P \geq 0 \) for COVID-19 negative. After setting the probability thresholds, we observed lower mispredictions and improved accuracy on the train-split dataset. Therefore, we used these thresholds on both train-split and test-split datasets and tabulated the count of positives and negatives in each threshold in Table 5. Using the thresholds on calibrated RF, overall, all samples had no false negatives and only two false positives. The false positives could be an experimental error in sample preparation or from the PCR analysis ground truth.

5.3. Random Forest Interpretation Using Explainable AI

To instill trust in the model’s predictions, we interpret the RF model both globally and locally using X-AI techniques. While global interpretation explains the importance of each feature in overall model prediction, local interpretation provides reasoning on a single instance level. For global interpretation, analyzing the importance of features determined by a trained model gives information about its prediction process. In this section, we leverage multiple techniques to investigate the feature importance of our trained RF model.

5.3.1. Impurity-based feature importance (IFI)

First, we interpreted the trained RF classifier with a model-specific IFI technique. This technique computes the feature importance as the average decrease in Gini impurity (Ceriani & Verme, 2012) from a feature node to its child nodes. Generally, except for the leaf nodes at the bottom of the tree, all other nodes have a threshold condition placed on a feature whose decision splits the node into two child nodes. The drop in impurity in this split is then calculated as the difference of Gini impurity between the node and the sum of its child nodes, weighted by the number of samples in each node. The final feature importance is the mean of drop in impurity at every prevailing feature node across the forest. Following these procedures, we calculated the IFI for all 10 features as shown in panel A of Figure 6 with the drop in Gini impurity on the y-axis and features on the x-axis. Although this method gives a general understanding of feature importance, one disadvantage is the inflation of importance for features with high cardinality in the forest. Therefore, we investigated other feature-importance algorithms to validate the results.

5.3.2. Permutation feature importance (PFI)

The PFI technique (Breiman, 2001) calculates the drop in performance when the trained RF algorithm is validated with permuted feature values which essentially makes the feature a noise. This method analyzes one feature at a time by permuting their values to break the association between the feature and the outcome. Clearly, the permutation of important features results in a larger performance drop. Therefore, the feature importance of using PFI is the drop in permuted feature accuracy from the base accuracy (where every feature was used in its original form). Bias from random value assignment to feature values was prevented by permuting values and averaging the drop in accuracy across

**Figure 5:** Probability of each sample (on the x-axis) to test COVID-19 positive (on the y-axis) for un-calibrated RF (on the left) and calibrated RF (on the right). The true labels for each sample are color-coded in blue and red for COVID-19 negative and positive tests, respectively. The grey triangles show the anomalies identified by the isolation forest.
1000 iterations. However, this algorithm fails to capture feature interaction effects as it permutes only one feature at a time. Panel B in Figure 6 shows the permutation importance scores for the RF classifier with overall feature importance printed on the y-axis and the features on the x-axis.

5.3.3. Shapely additive explanations (SHAP)

Lundberg & Lee (2017) formalized the SHAP algorithm as an additive feature attribution technique that unifies concepts from local model-agnostic explanations (LIME) (Ribeiro et al., 2016) and Shapely values (Shapley, 1953). In general, linear models are easy to interpret as the feature importance of the model is reflected by the model’s coefficients. However, due to the complexity of the RF algorithm, it cannot be interpreted as a simple linear model. The output of a single sample from RF, on the other hand, could be explained by producing some neighboring samples and then fitting a linear model locally on the newly obtained neighborhood samples. The LIME algorithm follows this approach: it generates a local point of view per sample by applying perturbations to its feature values and generating some artificial data in the neighborhood of a sample. LIME then fits a surrogate model on the artificial data and presents the coefficients of the surrogate model as the model interpretation. The downside of LIME is that it generates artificial samples using a heuristic kernel function. However, artificially replicating clinical data is difficult, and the dependence on kernel function heuristics to generate artificial samples is unreliable. SHAP adopts LIME procedures and bridges its limitations with the Shapely values, which borrow principles from cooperative game theory. Shapely values rationalize the distribution of credit or payoff among the players in a cooperative game by inspecting each player’s contribution individually and collaboratively as a team (Shapley, 1953).

In theory, the feature attributions assigned by our trained RF classifier \( f \) to the features \( R_i \) (where \( i \in (0, 9) \) for 10 ratio features) in a given data sample \( x \) is computed as \( \phi_R(f, x) \) using Equation (3). This equation facilitates SHAP to compute a unique optimal solution for feature attribution by combining individual as well as team efforts. This can be understood in light of PFI, where we extracted feature attributions by masking each feature at a time and computing the drop in performance of the classifier. PFI captures individual contributions but fails to grasp collaborative contributions. The more intuitive approach of SHAP allows for collaborative contributions in conjunction with individual contributions to be considered by iterating over all possible subsets \( S \) in our feature space having \( F \) number of features excluding feature \( R_i \) (where \( F = 10 \)) as \( S \subseteq F \setminus \{ R_i \} \). In each subset, the marginal contribution \( C \) of feature \( R_i \) is extracted as the difference in the outcome of the function \( f \) with and without \( R_i \) in the subset as shown in Equation (1) where \( f_{S \cup \{R_i\}} \) and \( f_S \) are two retrained functions with and without \( R_i \) respectively that are marginalized over features absent in the subset. The resultant sum of \( C \) is weighted with \( W \) (refer to Equation (2)) whose numerator is the product of the different number of ways a subset can be formed \( (S!) \) and the number of ways to choose features excluding the features in a subset \( (|F| - |S|-1)! ) \). The numerator is normalized by \( F! \), the total possible ways of choosing the features. Finally, for each feature in \( x \), Equation 3 computes the SHAP values through iterations over all possible subsets \( S \). The magnitude of final SHAP values shows the contribution of a feature to the outcome of the classifier. In this sense, the positive or negative SHAP value indicates the feature’s support towards the COVID-19 positive or negative decision respectively.

\[
C = \frac{f_{S \cup \{R_i\}}(x_{S \cup \{R_i\}}) - f_S(x_S)}{|F|!} 
\]

\[
W = \frac{|S|!(|F| - |S| - 1)!}{|F|!} 
\]

\[
\phi_R(f, x) = \sum_{S \subseteq F \setminus \{R_i\}} W \times C 
\]

Since the idea behind Shapely values is that there exists only one solution to allocate the attributions, the results are more reliable than previous methods such as LIME (Ribeiro et al., 2016). In addition to this, SHAP has other desirable axiomatic properties. The local accuracy property entails that the sum of all feature attributions of a model should approximate the original model meaning \( f(x) = \phi_{R_1} + \sum_{i=0}^{n} \phi_{R_i} \) where \( R_i \) is the feature attribution when all features are toggled off. The missingness property ensures that the features toggled off do not impact the model’s output. Lastly, the consistency property ensures for any feature \( R_i \), \( \phi_{R_i} \) remains consistent with respect to the impact of feature values of any retrained model. For example, in two functions \( (f_1, f_2) \) trained on different subsets including the feature \( R_i \), if the feature values of \( R_i \) have larger impact on \( f_1 \) than \( f_2 \), then \( \phi_{R_1}(f_1, x) \geq \phi_{R_1}(f_2, x) \).

We then combine Equations (1), (2), and (3) and compute SHAP values for all 132 data points excluding 20 anomalies where each data point has having 10 ratio features. The result is a SHAP value matrix of dimensions \( [132 \times 10] \) that will aid in the local and global model interpretation of our trained RF model. For global model interpretation, we extract the magnitude of feature importances in two steps: 1) take the absolute SHAP values, which nullify the effect of direction, and 2) compute the mean of absolute SHAP values across all samples. This results in condensing the \( [132 \times 10] \) matrix to a single array of 10 feature importances (as shown in panel C of Figure 6). By comparing all 3 feature importance algorithms in Table 7, it is conclusive that the two most important features are \( R_8 \) and \( R_9 \) while there was some variation among the least important features. In addition to global (all-sample) explanations, we present local (per-sample) explanations using the SHAP value matrix as shown in Figure 9. The x-axis represents the SHAP value per sample, and features are numbered on the y-axis. We encoded the feature values in color ranging from blue (signifying...
Figure 6: Feature importance for RF classifier using three feature interpretation techniques: Impurity-based feature importance (panel A, top), Permutation Feature Importance (panel B, middle), and Shapely Additive Explanations (panel C, bottom). The y-axis shows the feature importance for each feature shown on the x-axis.

small value) to red (signifying large value) and assigned each feature to a quartile based on its value, as shown at the origin of each bar. The magnitude of each bar explains the impact of each feature on the model’s output and thus allowing us to interpret them as feature importance. We provide more details on global and local explanations in sections 6.2.1 and 6.2.2, respectively.

| Model          | Accuracy | Precision | Recall |
|----------------|----------|-----------|--------|
| RF             | 84.78%   | 0.83      | 0.84   |
| RF + AD        | 87.8%    | 0.84      | 0.89   |
| Cal RF + AD + PT | 94.12%   | 0.93      | 0.93   |

AD: Anomaly detection, PT: Probability threshold
Cal RF: Calibrated random forest

Table 6

Ablation experiment: add each component of our AI framework and observe their influence on the train-test split performance.

6. RESULTS

6.1. Results of Random Forest Classifier

We trained the RF algorithm with the hyperparameters extracted through an exhaustive grid search, shown in Table 4. Then, we removed outliers and employed a total of 91 samples in training and 41 in testing the model. We then calibrated the trained RF algorithm using Platt scaling and set a threshold on predicted probability for COVID-19 diagnosis at 0.7 and 0.25 for COVID-19 positive and COVID-19 negative diagnosis, respectively. The rationale for setting a threshold on probability is that samples on the border of diagnosis (i.e., probability == 0.5) may require special attention or re-testing. All in all, 28 of the samples had low confidence, and the remaining 104 had high confidence in probability. We conducted an ablation experiment by adding each component (i.e., anomaly detection, calibration, and probability threshold) to the baseline RF model. The results are shown in Table 6. First, we observed that the baseline RF model trained and tested with the train-test split on the whole dataset had accuracy, precision, and recall as 84.78%, 0.83, and 0.84, respectively. Next, adding anomaly detection and filtering 20 anomalous data points, we saw a slight improvement in performance with accuracy, precision, and recall to 87.8%, 0.84, and 0.89, respectively. Lastly, after calibrating the RF model and setting the probability threshold on the clean data without anomalies, the accuracy improved significantly to 94.12% and precision and recall both to 0.93. Because of this improved performance, we used the calibrated RF with anomaly detection and probability thresholds in our explainable-AI framework.

Since our dataset had more negative samples than positive samples, it was crucial to give equal weight to model metrics for both positive and negative categories. Therefore, we evaluated precision, recall, and F1 score metrics on both the test and train datasets and reported the macro average of the metrics on positive and negative COVID-19 samples. As shown in the evaluation report in Figure 7, the train-split had a high precision, recall, and F1-score of 0.98, 0.97, 0.97, respectively. On the other hand, precision, recall, and F1-scores were 0.93, 0.93, and 0.93, respectively, for the test-split. There was 1 false positive, 1 false negative, and 32 true predictions for the test-split.

Since our dataset had more negative samples than positive samples, it was crucial to give equal weight to model metrics for both positive and negative categories. Therefore, we evaluated precision, recall, and F1 score metrics on both the test and train datasets and reported the macro average of the metrics on positive and negative COVID-19 samples. As shown in the evaluation report in Figure 7, the train-split had a high precision, recall, and F1-score of 0.98, 0.97, 0.97, respectively. On the other hand, precision, recall, and F1-scores were 0.93, 0.93, and 0.93, respectively, for the test-split. There was 1 false positive, 1 false negative, and 32 true predictions for the test-split. In the train-split, there
were 2 false negatives and a total of 68 correct predictions. Although these metrics portray the high efficiency of the trained model in COVID-19 testing, it still does not provide a means to understand its prediction mechanisms. This motivated us to further dissect the trained RF model and understand its prediction mechanisms using X-AI techniques.

### 6.2. Model Interpretation of Random Forest Classifier

The outcome of the RF model is more complex than a simple decision tree. In decision trees, we have a single decision path for each prediction, making the model self-interpretable. However, RF coalesces the outcome of many unique decision trees, meaning for each sample, there are a multitude of decision paths; one for each tree traversal which makes model interpretation complex. Therefore, we simplify the interpretation of RF by presenting the rationale behind predictions in the form of global (all sample basis) and local (per sample basis) interpretations. These interpretations help domain experts and clinicians handling COVID-19 testing understand the classifier’s operations and gain rich diagnostic information. First, the idea of the global interpretation of a model goes beyond the analysis of feature importance. It can elaborate on how the distribution of each feature affects the model prediction. Second, local interpretation operates on one sample at a time and delineates each feature’s value and contribution to the outcome of RF.

#### 6.2.1. Global interpretation

While computing feature importance with different algorithms, we found similar feature ranking: \( R_9 \) and \( R_5 \) were consistent among the two features (as shown in Table 7). It is fairly common to observe slight differences in the feature importance across the three algorithms due to underlying procedural differences and the shortcomings of techniques highlighted in their respective sections. However, on the whole, these rankings reveal that most feature importances assigned by the three algorithms are fairly consistent. The consistent nature of feature attribution shows that our RF algorithm can be understood from different perspectives of the X-AI algorithms. Although these are important findings, they do not explain how the model prediction correlates with the feature values. While IFI and PFI algorithms can generate feature rankings, they do not provide enough information to explore the interactions between features. Therefore, we extended our X-AI approach to the SHAP algorithm, which can analyze feature interactions, data distribution, and the model’s dependency on it.

We utilized the SHAP algorithm to produce a violin plot, which provides a fine-grained view of SHAP values (feature importance) and feature value distribution as shown in Figure 8. This plot illustrates the trends of SHAP values in conjunction with feature values by plotting the SHAP value matrix (as described in section 5.3.3) without aggregation across all test-split samples. In addition, the plot encodes the density of the samples with a certain SHAP value as the width of the line. To interpret these plots, the higher the SHAP value, the greater the importance; the positive SHAP values support positive COVID-19 prediction, negative values support negative output, and the feature values are represented on a blue-red color-bar where deep blue and deep red correspond to low and high feature values. For example, the feature \( R_9 \) has SHAP values extending up to +0.495 towards a positive COVID-19 decision, where the feature value is low. For a negative test, mostly high values of \( R_9 \) contributed towards a negative COVID-19 prediction. On
the contrary, for feature $R_6$, lower and higher feature values contribute towards COVID-19 negative and positive tests, respectively. Features $R_6$ and $R_7$ follow similar patterns as $R_3$ while $R_0$ and $R_1$ have similar patterns of $R_0$.

![Diagram](image)

**Figure 8: Violin plot for global interpretation of calibrated RF classifier using SHAP values.**

The violin plot provides an approximate idea of any sample’s model output based on the feature distribution. This can be immensely helpful to cross-examine the model if the model prediction is not following the trend of the violin plot. For example, suppose the model is predicting COVID-19 positive for a sample where the $R_0$ feature value is very high. In that case, we can easily identify that the model is not following the trend, and hence, it needs further validation. It also shows any outliers in the dataset, for instance, higher feature values of $R_8$ (colored in deep red) have a large SHAP value. However, the width of the line is very thin which indicates that the number of samples is potentially lacking or is the case of an outlier. This kind of information is critical to instill trust in users as it allows them to question outputs from the model.

### 6.2.2. Local interpretation

We present a comprehensive analysis of the RF classifier’s interpretation across various prediction scenarios, including a true positive (TP), a false positive (FP), a true negative (TN), and a false negative (FN) (as shown in Figure 9 and Table 8). In Figure 9, each bar is encoded with a color gradient ranging from shades of blue to red, representing the lowest and highest feature values, respectively. Similar to the global SHAP plots, the feature importance (or the SHAP values) are shown along the x-axis. A positive value signifies support for COVID-19 positive, while a negative for COVID-19 negative classification. In addition to SHAP values, the feature values can be employed to visualize their impact on a COVID-19 positive or negative decision. After analyzing several local interpretation plots, we found that the feature value was easier to interpret using color bars than discrete feature values. Therefore, we infer the feature values as low-range (deep blue), high-range (deep red), and mid-range (between deep blue and deep red). The feature values help us see how each sample compares against the rest of the samples in the dataset with the aid of global interpretations in Figure 8. For instance, the feature $R_9$ in panel A is deep blue, indicating a low feature. Now, in Figure 8, the distributions of $R_9$ feature value at the lower end (i.e., deep blue) and the higher end of the spectrum (i.e., deep red) are relatively dense compared to the mid-region. As discussed in Section 6.2.1, the density of samples in the dataset at any given SHAP value can be realized by the width of the marker in the violin plot. The local interpretation enables us to comprehend how the classifier makes decisions for individual instances by accounting for the feature interactions, shedding light on the model’s behavior, and enhancing our trust in its predictions.

### 6.2.3. Explainable AI Framework

In this section, we present our X-AI framework that compiles model prediction, feature importance, and global and local model interpretation in a systematic framework as illustrated in Figure 10. We organized our X-AI framework in four stages, each designed to systematically interpret the decisions made by our Random Forest (RF) classifier and reach a final conclusion. In each stage, we focus on a specific aspect of interpretation. First, we check the alignment of the local interpretation with the feature importance. Second, we assess the local interpretation and contrast it with the global interpretation to ensure that each feature is consistent with the global interpretation. Lastly, in the third stage, we evaluate the consistency of feature importance in the local interpretation. Following this approach, our framework provides clear, concise, and easy-to-understand insights into the model’s predictions, facilitating a comprehensive analysis of the underlying decision-making process. We elaborate on the relevance and properties of each stage below.

- **Global feature importance** In this stage, we compute the feature importance scores and generate global interpretation scores. The feature importance and global model interpretation results will be used in stage-3-check-1 and stage-3-check2, respectively.

- **Model prediction (Stage-1)** This stage provides the binary COVID-19 outcome from our RF classifier.

- **Model probability (Stage-2)** Going beyond the prediction, this stage shows us the probability of the
RF model. Here, if the probability is not within the threshold, then it is treated as an invalid test due to low confidence.

- **Feature importance to local feature importance agreement (stage-3-check-1)**
  In this stage, we investigate the local feature importance and compare it with the SHAP feature importance displayed in Figure 6 panel C. From the global feature importance analysis, we identified the most important features as \( R_9, R_8, R_7, R_5, \) and \( R_0 \). By examining whether these features exhibit similar importance in individual samples, we can assess if the sample-based feature importance aligns with the model’s feature importance derived from the training dataset.

- **Global and local feature importance and value similarity (stage-3-check-2)**
  In this step, we analyze the similarity between global and local interpretations of feature importance and feature value. To do this, we compare individual samples from the local interpretation to the global interpretation presented in Figure 8. Recall that the violin plot visualizes the distribution of feature values, with a wider marker indicating more samples and a narrower line representing fewer samples. Samples in narrow regions could be outliers based on feature value or receive special attention from the model, which could lead to undesirable results. For instance, a small change in feature value may result in a significant shift in model prediction. To ensure that local interpretations are consistent with global trends, we verify whether

---

**Figure 9:** Local explanations generated using SHAP algorithm for six records for COVID-19 positive predictions (panels A, B, and C) and COVID-19 negative predictions (panels D, E, and F). Panels A and D show highly confident predictions, panels B and E show low confident predictions, and panels C and F show false positive and false negative predictions, respectively.
Check-1 Are the SHAP feature importances correlated with local feature importance? (For example, check if $R_5$, $R_8$, $R_7$, $R_9$, $R_5$ are also important in local importance.)

Check-2 Do the feature importance and feature value on local interpretation resemble the majority in the global interpretation in Figure 8?

Check-3 Are the feature importances in local interpretation consistent in their support for COVID-19 outcome? (i.e., all features support COVID-19 negative or positive outcomes.)

Table 8
An explainable AI framework for COVID-19 diagnosis decision making.

| Index | True label | Stage-1 | Stage-2 | Stage-3 | Stage-4 |
|-------|------------|---------|---------|---------|---------|
|       | Model prediction | Inspect probability | Turn to feature interpretation | Final decision |
|       |             |         | Check-1 | Check-2 | Check-3 |
| Panel-A | Positive | Positive | 0.99 | ✓ | ✓ | ✓ | Valid test |
| Panel-B | Positive | Positive | 0.67 | ✓ |✗ |✗ | Re-inspect |
| Panel-C | Positive | Negative | 0.13 | ✓ | ✓ | ✓ | Re-inspect/ valid test |
| Panel-D | Negative | Negative | 0.07 | ✓ | ✓ | ✓ | Valid test |
| Panel-E | Negative | Negative | 0.41 | ✓ | ✓ | ✓ | Re-inspect |
| Panel-F | Negative | Positive | 0.92 | ✓ | ✓ | ✓ | Re-inspect |

Figure 10: Our Explainable AI framework.

The samples lie in the denser regions of the violin plot, confirming their adherence to the majority. By comparing local and global interpretations in this manner, we can assess whether the model’s behavior is consistent across different scales and better understand its decision-making process. Through this analysis, we aim to identify potential issues or inconsistencies in the model’s interpretation of feature importance and feature value, which could be critical for improving model performance, addressing potential biases, and enhancing overall interpretability.

Consistency in local feature importance support for COVID-19 outcomes (stage-3–check-3)
In this check, we evaluate the consistency of feature importance in local interpretations concerning their support for COVID-19 outcomes, i.e., positive or negative. The objective is to determine if the contributions of features in individual samples consistently favor either a positive or negative COVID-19 outcome. Ensuring such consistency is vital for the interpretability and reliability of the explainable AI framework, as it helps to establish that the model’s local predictions are based on coherent and meaningful feature contributions, thereby facilitating a better understanding of the model’s decision-making process.

6.2.4. Case Study
We first analyze a true positive prediction by the RF algorithm, with a probability of $P = 0.99$ (Figure 9, panel A). This sample clearly passes stage-2 since the probability exceeds the threshold of > 0.7. In stage-3, the sample passes check-1 because the most important local features, $R_9$ and $R_8$, align with the most important features in the global feature importance. In check-2, we compare the feature values in relation to their SHAP values and contrast them with the global model interpretation presented in the violin plot 8. We observe that features $R_5$, $R_3$, $R_1$, and $R_2$ have large feature values, while others possess smaller values. Comparing the feature value color to the dense region of the violin plot,
we find that it falls within the majority distribution, thus passing check-2. Finally, in check-3, we examine whether all SHAP values are consistent (i.e., uniformly supporting a positive or negative decision). Although one important feature, $R_1$, supports a negative decision, all other features display consistency, leading the sample to pass check-3 and be classified as a valid test. Next, in panel D, we show a confident prediction for a true negative case. Here, the sample passes stage-2, and all checks in stage-3 as the SHAP values align with feature importance, resemble the major distribution, and are consistent, similar to panel A.

Two cases with low confidence in the prediction for COVID-19 positive and negative (i.e., the probability is close to 0.5) are shown in Figure 9, panels B and E. Panel B exhibits a true positive prediction similar to panel A; however, panel B has lower confidence in the prediction with a predicted probability of $P = 0.67$ rendering it an invalid test. In stage-3, panel B passes check-1 as the features $R_5$, $R_8$, and $R_9$ have high SHAP values as the global feature importance. However, panel B fails check-2 since the features $R_3$-$R_6$ do not fall in the dense regions for positive SHAP values in the global importance. Lastly, panel B does not pass check-3, as well as it has four features supporting each of the positive and negative classes, and the remaining two have close to 0 SHAP values. In our second example for low confidence prediction ($P = 0.41$) in panel E, we see that it passes check-1 as the blue color in feature value can be found in dense regions of the violin plot. However, it fails check-2 and check-3 for the same reasons as panel B. Since both panel B and panel E failed stage 3, they need a re-inspection or re-testing.

Lastly, we analyze two incorrect predictions in our test set: a false negative (panel C) with a probability of $P = 0.13$ and a false positive (panel F) with a probability of $P = 0.92$. In panel C, even though the sample belongs to the positive class, all SHAP values support a negative outcome. Notably, the SHAP values of panel C match those of panel D (a true negative prediction). This similarity arises because the positive sample has feature values closely resembling the negative sample, leading to a similar prediction for panel C as for panel D. Panel C does not pass check-1 since only features $R_9$ and $R_8$ are important, while the others have minimal contributions. The result’s outcome may necessitate considering patient symptom information to determine whether a re-test is required or if the result can be deemed valid. In panel F, the feature importance magnitudes match the global feature importance; however, the most important feature, $R_8$, displays an inversion, while the other features support a positive prediction. This indicates that panel F passes check-1 and check-2 but fails check-3.

7. DISCUSSION

While analyzing our results through different interpretation techniques, we observed quite remarkably that all three methods of global feature interpretation consistently resulted in the most important features as $R_8$ and $R_9$ (Table 7). The meaning of these features is of particular interest because of the biomarker ranges making up the features, $AUC_{B_1}/AUC_{B_9}$ for $R_8$ and $AUC_{B_3}/AUC_{B_7}$ for $R_9$. First, the features $R_8$ and $R_9$ are virus-to-host features, which may represent the interaction between host immune response and viral antigen, as described in Chivite et al. (2021). Second, the viral protein fragment corresponding to $B_3$ is tentatively suggested to be the S1 subunit of the viral spike protein (see Table 1). S1 is a key viral biomarker in the mass spectrometry because the observed m/z ratio does not overlap significantly with other proteins or fragments from the gargle sample. Furthermore, because our model uses SHAP values, it was particularly remarkable that $R_9$ was the most important feature, as it contains $B_3$. Under a similar interpretation of our model, $R_8$ was deemed to be the least important feature, which contains $B_4$ and $B_5$, a host-to-host protein ratio. As no viral protein is represented in $R_9$, it follows that such a feature would not be useful in the diagnosis of COVID-19. It should be noted once again that each biomarker has been assigned a putative identity based on preliminary studies only and have yet to be unambiguously characterized.

Turning to local feature interpretations, these could offer a powerful, comprehensive diagnostic view in contrast to a binary decision by RT-PCR or an algorithm with percent confidence. The local explanations generated using SHAP values show the contribution of each feature in each sample. This could allow domain experts and health officials administering such a test to make more informed decisions based on the putative host-to-viral or viral-to-host ratios for case-by-case bases, as seen in Figure 9. If the algorithm decides positive or negative, the local feature explanation will provide additional information as to why the test output its decision. Although data augmentation techniques could enrich the dataset with more artificial samples, we avoided doing this because it could have made the dataset noisy by adding false patterns. Nonetheless, not only in this diagnostic approach is the viral aspect of COVID-19 considered (as is only considered in PCR-based tests), but the human immune response towards the SARS-CoV-2 virus is also given due consideration. The results show that machine learning algorithms (e.g., RF) coupled with the utility of X-AI interpretability could be implemented into healthcare to address the limitations of black box systems and improve the diagnostic process, from test output to informed care.

In summary, from training machine learning algorithms to explaining the trained models, each aspect may be particularly interesting for different reasons. For example, during the AI-model building stages, a closer view of calibration, probabilities, confusion matrices, precision, and recall values for the training and testing datasets can prevent overfitting and bias. Following that, the interpretation of a model on a global and local level can have several applications. First, from a machine learning perspective, feature importance can help identify spurious correlations between features and diagnoses and monitor model drift. Second, biochemistry researchers can learn and investigate the relationships between

: Preprint submitted to Elsevier
feature value to its importance in the outcome of COVID-19 diagnosis. Finally, in a clinical setting, the local model interpretations, along with the probability scores, can aid in determining whether a test was properly executed or if a follow-up test was required.

In the case of a future pandemic, MALDI-ToF coupled with AI can be used as an alternative or in-addition-to-testing platform for diagnosing viral infections. MALDI-ToF MS could be advantageous in clinical virology for the option of automation, a low cost-to-efficiency ratio, rapid identification, high throughput testing, high accuracy, and reliability. Even though the cost of a MALDI-ToF instrument is higher than the materials and instrumentation for RT-PCR, purchasing a MALDI-ToF mass spectrometer is a one-time investment. Furthermore, the cost-per-test for RT-PCR testing is significantly higher. It has been estimated by Deulofeu et al. (2021) that the cost for a MALDI-ToF test is at least 75% lower than an RT-PCR test. A MALDI-ToF-based testing platform could allow scientists to develop and validate novel methods for the diagnosis of diseases. As we have seen in this pandemic, the complete reliance on the gold standard testing technique limits the ability to perform early mass testing due to supply limitations. In addition, although previous research has shown some good results incorporating AI algorithms, in this study, we show adding an additional layer of X-AI and designing a three-stage outcome can improve efficiency and generate explanations from different perspectives.

8. CONCLUSION

In this work, we have designed an AI-based COVID-19 diagnostic method using the MALDI-ToF MS protein profiles of human gargle samples. After thoroughly validating various machine learning algorithms, we found the performance of RF to be the best among others. We evaluated the RF classifier on a 70%–30% train-test-split strategy where the accuracy on the test dataset was 94.12%. We further utilized the concepts of X-AI and interpreted the RF algorithm using SHAP and feature importance techniques such as permutation feature importance and impurity-based feature importance to find the most important features. Using these interpretation models, we showed global and local explanations of the model’s decision on COVID testing. Explaining the rationale behind the decision improve your trustworthiness of our approach by enabling domain experts and medical practitioners to understand the mechanisms of the black-box AI models when used on a real-time basis. Additionally, this methodology may allow for more detailed information about the life cycle of viral infection along with the COVID-19 status ascertained from the MALDI-ToF MS data.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the publication of the article.

Acknowledgements

This work was supported by Northern Illinois University’s Molecular Analysis Core which was established with support from Shimadzu Scientific Instruments.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: a system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (osdi 16) (pp. 265–283).

Abraham, J. E., Maranian, M. J., Spiteri, L., Russell, R., Ingle, S., Lucarini, C., Earl, H. M., Pharoah, P. P., Dunning, A. M., & Caldas, C. (2012). Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. BMC Medical Genomics, 5, 1–6.

Alves, M. A., Castro, G. Z., Oliveira, B. A. S., Ferreira, L. A., Ramírez, J. A., Silva, R., & Guimarães, F. G. (2021). Explaining machine learning based diagnosis of covid-19 from routine blood tests with decision trees and criteria graphs. Computers in Biology and Medicine, 132, 104335.

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks. Computers in biology and medicine, 121, 103795.

Basu, A., Sheikh, K. H., Cuevas, E., & Sarkar, R. (2022). Covid-19 detection from ct scans using a two-stage framework. Expert Systems with Applications, 193, 116377.

Benjamini, Y., & Hochberg, Y. (1995, January). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300.

Breiman, L. (2001, October). Random forests. Machine Learning, 45(1), 5–32.

Carter, L. J., Garner, L. V., Smoot, J. W., Li, Y., Zhou, Q., Saveson, C. J., Sasso, J. M., Gregg, A. C., Soares, D. J., Beskid, T. R., Jervey, S. R., & Liu, C. (2020, April). Assay techniques and test development for COVID-19 diagnosis. ACS Central Science, 6(5), 591–605.

Ceriani, L., & Verme, P. (2012, September). The origins of the gini index: extracts from variabilita em mutabilita (1912) by corrado gini. The Journal of Economic Inequality, 10(3), 421–443.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Chivite, P., LaCasse, Z., Seethi, V. D. R., Bharti, P., Bland, J., Kadkol, S. S., & Gaillard, E. R. (2021, August). MALDI-ToF protein profiling as a potential rapid diagnostic platform for COVID-19. Journal of Mass Spectrometry and Advances in the Clinical Lab, 21, 31–41.

Costa, M. M., Martin, H., Estellon, B., Dupé, F.-X., Saby, F., Benoit, N., Tissot-Dupont, H., Million, M., Pradines, B., Grangeaud, S., & Almeras, L. (2022, January). Exploratory study on application of MALDI-TOF-MS to detect SARS-CoV-2 infection in human saliva. Journal of Clinical Medicine, 11(2), 295.

de Almeida, C. M., Motta, L. C., Follí, G. S., Marcariani, W. D., Costa, C. A., Vilêla, A. C., Barauna, V. G., Martin, F. L., Singh, M. N., Campos, L. C., Costa, N. L., Vassallo, P. F., Chaves, A. R., Endringer, D. C., Mill, J. G., Filgueiras, P. R., & Ramão, W. (2022). MALDI (+) FT-ICR Mass Spectrometry (MS) combined with machine learning toward saliva-based diagnostic screening for COVID-19. Journal of Proteome Research, 21(8), 1868–1875.

Deulofeu, M., Garcia-Cuesta, E., Peña-Méndez, E. M., Conde, J. E., Jiménez-Romero, O., Verdú, E., Serrano, M. T., Salvador, V., & Boadas-Vello, P. (2021, April). Detection of SARS-CoV-2 infection in human nasopharyngeal samples by combining MALDI-ToF-MS and artificial intelligence. Frontiers in Medicine, 8, 398.

Esser, D., Álvez-Llamas, G., De Vries, M. P., Weening, D., Vonk, R. J., & Roelofs, H. (2008). Sample stability and protein composition of
Axima performance - a highly sensitive Shimadzu. (2022, Accessed: March).

Rahman, M. A., Hossain, M. S., Alrajeh, N. A., & Gupta, B. (2021).

Shapley, L. S. (1953, March). A value for n-person games.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Rish, I. (2001). An empirical study of the naive bayes classifier. In Ijcai 2001 workshop on empirical methods in artificial intelligence (Vol. 3, pp. 41–46).

Rocca, M. F., Zintgraf, I. C., Datiero, M. E., Santos, L. S., Ledesma, M., Rostami, M., & Oussalah, M. (2022). A novel explainable COVID-19 diagnosis method by integration of feature selection with random forest. Informatics in Medicine Unlocked, 30, 100941.

Safavian, S. R., & Landgrebe, D. (1991, May). A review of interpretability methods. Computers in Biology and Medicine, 140, 105111.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the ieee international conference on computer vision (pp. 618–626).

Shapley, L. S. (1953, March). A value for n-person games. Contributions to the Theory of Games 2, 302(28), 307–317.

Sharma, N., Saba, L., Khanna, N. N., Kalra, M. K., Fouda, M. M., & Suri, J. S. (2022). Segmentation-based classification deep learning model embedded with explainable ai for covid-19 detection in chest x-ray scans. Diagnostics, 12(9), 2132.

Shimadzu. (2022, Accessed: March). Axima performance - a highly sensitive Shimadzu. (2022, Accessed: March).

Sohrabi, C., Afsari, Z., O’neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., & Agha, R. (2020, April). World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery, 76, 71–76.

Spick, M., Lewis, H. M., Wilde, M. J., Hopye, C., Huggett, J., & Bailey, M. J. (2022, January). Systematic review with meta-analysis of diagnostic test accuracy for COVID-19 by mass spectrometry. Metabolism, 126, 154922.

Stähle, L., & Wold, S. (1987, July). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. Journal of Chemometrics, 1(3), 185–196.

Tallarida, R. J., & Murray, R. B. (1987). Area under a curve: trapezoidal and Simpson’s rules. Manual of Pharmacologic Calculations, 77–81.

Thimoteo, L. M., Velasco, M. M., Amaral, I., Figueiredo, K., Yokoyama, C. L., & Marques, E. (2022). Explainable artificial intelligence for COVID-19 diagnosis through blood test variables. Journal of Control, Automation and Electrical Systems, 33(2), 625–644.

Tibshirani, R. (1996, January). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Tideman, L. E., Migas, L. G., Djambazova, K. V., Patterson, N. H., Caprioli, R. M., Spragins, J. M., & Van de Plas, R. (2021). Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized shapley additive explanations. Analytica Chimica Acta, 1777, 338522.

Tolman, N. K., Howard, T., Walsh, R., Pepper, J., Loegering, J., Phinney, B., Salamini, M. R., & Rishadi, H. H. (2021, April). Novel application of automated machine learning with MALDI-ToF-MS for rapid high-throughput screening of COVID-19: a proof of concept. Scientific Reports, 11(1), 1–10.

Ullah, F., Moon, J., Naeem, H., & Jabbar, S. (2022). Explainable artificial intelligence approach in combating real-time surveillance of covid19 pandemic from ct scan and x-ray images using ensemble model. The Journal of Supercomputing, 1–26.

Valera, E., Jankelow, A., Lim, J., Kindratenko, V., Ganguli, A., White, K., Kumar, J., & Bashir, R. (2021, May). COVID-19 point-of-care diagnostics: present and future. ACS Nano, 15(5), 7899–7906.

Vinod, D. N., Jeyavadhavan, B. R., Zanguru, A. M., & Prabaharan, S. (2021). Fully automated unified prognostic diagnosis of covid-19 chest x-ray scan images using deep covix-net model. Computers in Biology and Medicine, 136, 104729.

Virtanen, P. G., Pommers, O. R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H. D., Pedregosa, F., van Mulbregt, P., Vijayakumar, A., Bardelli, A. P., Rothenberg, A., Hilboll, A., Kloekner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A. H., Hagen, D. R., Pasechnik, D. V., Olivi, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. J., Allen, G. E. H., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silvert, J., Webber, J. T., Slivice, J., Nothman, J., Buchner, J., Kulick, J., Schönböger, J. L., de Miranda Cardoso, J. V., Reimer, I., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y., & Contributors, S. (2020). Scipy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, 17(3), 261–272.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-cam: Score-weighted visual explanations for
convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 24–25).

Wen, C., Liu, S., Liu, S., Heidari, A. A., Hijji, M., Zarco, C., & Muhammad, K. (2022). ACSN: Attention capsule sampling network for diagnosing COVID-19 based on chest CT scans. *Computers in Biology and Medicine*, 106338.

(WHO) World Health Organization. (2023, Accessed: March). *WHO COVID-19 Dashboard*. Retrieved from https://covid19.who.int/

Wold, S., Esbensen, K., & Geladi, P. (1987, August). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52.

Xie, Y. R., Castro, D. C., Bell, S. E., Rubakhin, S. S., & Sweedler, J. V. (2020). Single-cell classification using mass spectrometry through interpretable machine learning. *Analytical chemistry*, 92(13), 9338–9347.

Xu, Y., Lam, H.-K., Jia, G., Jiang, J., Liao, J., & Bao, X. (2023). Improving covid-19 ct classification of cnns by learning parameter-efficient representation. *Computers in Biology and Medicine*, 152, 106417.

Yan, L., Yi, J., Huang, C., Zhang, J., Fu, S., Li, Z., Lyu, Q., Xu, Y., Wang, K., Yang, H., Ma, Q., Cui, X., Qiao, L., Sun, W., & Liao, P. (2021). Rapid detection of COVID-19 using MALDI-TOF-based serum peptidome profiling. *Analytical chemistry*, 93(11), 4782–4787.

Ye, Q., Xia, J., & Yang, G. (2021). Explainable AI for COVID-19 CT classifiers: An initial comparison study. In *2021 IEEE 34th international symposium on computer-based medical systems (cbms)* (pp. 521–526).