Machine Learning Classification Models for COVID-19 Test Prioritization in Brazil

Íris Viana dos Santos Santana, Andressa C. M. da Silveira, Álvaro Sobrinho, Lenardo Chaves e Silva, Leandro Dias da Silva, Danilo Freire de Souza Santos, Edmar Candeia, Angelo Perkusich

Submitted to: Journal of Medical Internet Research
on: January 25, 2021

Table of Contents

Machine Learning Classification Models for COVID-19 Test Prioritization in Brazil

Íris Viana dos Santos Santana[1]; Andressa C. M. da Silveira[2] MSc; Álvaro Sobrinho[3, 1] PhD; Lenardo Chaves e Silva[4] PhD; Leandro Dias da Silva[3] PhD; Danilo Freire de Souza Santos[2] PhD; Edmar Candeia[2] PhD; Angelo Perkusich[2] PhD

[1]Federal University of the Agreste of Pernambuco Garanhuns BR
[2]Federal University of Campina Grande Campina Grande BR
[3]Federal University of Alagoas Maceió BR
[4]Federal Rural University of the Semi-Arid Pau dos Ferros BR

Corresponding Author:
Álvaro Sobrinho PhD
Federal University of the Agreste of Pernambuco
Av. Bom Pastor, s/n - Boa Vista
Garanhuns
BR

Abstract

Background: controlling the COVID-19 outbreak in Brazil is considered a challenge of continental proportions due to the high population and urban density, weak implementation and maintenance of social distancing strategies, and limited testing capabilities.

Objective: to contribute to addressing such a challenge, we present the implementation and evaluation of supervised Machine Learning (ML) models to assist the COVID-19 detection in Brazil based on early-stage symptoms.

Methods: firstly, we conducted data preprocessing and applied the Chi-squared test in a Brazilian dataset, mainly composed of early-stage symptoms, to perform statistical analyses. Afterward, we implemented ML models using the Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Decision Tree (DT), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost) algorithms. We evaluated the ML models using precision, accuracy score, recall, the area under the curve, and the Friedman and Nemenyi tests. Based on the comparison, we grouped the top five ML models and measured feature importance.

Results: the MLP model presented the highest mean accuracy score, with more than 97.85%, when compared to GBM (> 97.39%), RF (> 97.36%), DT (> 97.07%), XGBoost (> 97.06%), KNN (> 95.14%), and SVM (> 94.27%). Based on the statistical comparison, we grouped MLP, GBM, DT, RF, and XGBoost, as the top five ML models, because the evaluation results are statistically indistinguishable. The ML models` importance of features used during predictions varies from gender, profession, fever, sore throat, dyspnea, olfactory disorder, cough, runny nose, taste disorder, and headache.

Conclusions: supervised ML models effectively assist the decision making in medical diagnosis and public administration (e.g., testing strategies), based on early-stage symptoms that do not require advanced and expensive exams.

(JMIR Preprints 25/01/2021:27293)
DOI: https://doi.org/10.2196/preprints.27293

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Machine Learning Classification Models for COVID-19 Test Prioritization in Brazil

Íris Viana dos Santos Santana[1], Andressa C. M. da Silveira[2], Álvaro Sobrinho[1,3*], Lenardo Chaves e Silva[4], Leandro Dias da Silva[3], Danilo Freire de Souza Santos[2], Edmar Candeia[2], and Angelo Perkusich[2]

Federal University of the Agreste of Pernambuco, Brazil[1]
Federal University of Campina Grande, Brazil[2]
Federal University of Alagoas, Brazil[3]
Federal Rural University of the Semiarid, Brazil[4]

Corresponding author: alvaro.alvares@ufape.edu.br*

Abstract

Background: controlling the COVID-19 outbreak in Brazil is a challenge of continental proportions due to the population's size and urban density, inefficient maintenance of social distancing and testing strategies, and limited availability of testing resources.
Objective: the purpose of this study is to effectively prioritize symptomatic patients for testing to assist the early COVID-19 detection in Brazil, addressing problems related to inefficient testing and control strategies.
Methods: raw data from 55,676 Brazilians were pre-processed, and the Chi-squared test was used to confirm the relevance of features: *Gender*, *Health Professional*, *Fever*, *Sore Throat*, *Dyspnea*, *Olfactory Disorders*, *Cough*, *Coryza*, *Taste Disorders*, and *Headache*. Classification models were implemented relying on pre-processed datasets, supervised learning, and the algorithms Multilayer Perceptron (MLP), Gradient Boosting Machine (GBM), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR). The models' performances were analyzed using 10-fold cross-validation, classification metrics, and the Friedman and Nemenyi statistical tests. The permutation feature importance method was applied for ranking the features used by the classification models with the highest performances.
Results: *Gender*, *Fever*, and *Dyspnea* are among the highest-ranked features used by classification models. The comparative analysis presents MLP, GBM, DT, RF, XGBoost, and SVM as the highest performance models with similar results. KNN and LR were outperformed by the other algorithms. Applying the easy interpretability as an additional comparison criterion, the DT was considered the most suitable model.
Conclusions: the DT classification model can effectively (e.g., mean accuracy ≥ 89.12%) assist the COVID-19 test prioritization in Brazil. The model can be applied to recommend the prioritizing of a symptomatic patient for COVID-19 testing.

Keywords: COVID-19; Test Prioritization; Classification Models; Medical Diagnosis

Introduction

Overview

In modern medical systems, healthcare professionals, managers, and governments use information and data analysis to make decisions [1]. Data is stored, enabling rapid access and sharing during the

diagnosis, monitoring, and treatment of patients. Therefore, there are propositions of e-Health and m-Health systems to assist healthcare professionals and policymakers with decision-making [2, 3]. Such systems are relevant to provide decision-support advice based on patients' data, helping healthcare professionals and policymakers address problems related to inefficient COVID-19 testing and control strategies (e.g., limited testing resources) in low and middle-income countries [4]. For example, people who live in low and middle-income, remote, and hard-to-reach settings are the most affected by precarious health care. Such a situation is even more critical in a pandemic scenario.

The COVID-19 is a disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [5]. In December 2019, the first cases of COVID-19 appeared in Wuhan, Hubei province, China [6]. Due to the high growth of COVID-19 confirmed cases worldwide, on January 30, 2020, the World Health Organization considered the COVID-19 outbreak a public health emergency of international importance [7].

Motivation and Problem Statement

As the number of COVID-19 confirmed cases continuously increases, healthcare professionals and policymakers need to define guidelines to prevent the disease, delaying the transmission rates. Such guidelines are relevant due to the high probability of collapse in health services and shortages of medical supplies (e.g., testing resources) [8]. Confirmation of the first COVID-19 in Brazil was in March 2020, and since then, there has been an upward trend in confirmed cases and deaths. Unfortunately, the Brazilian government has reported more than 11 million cases, with more than 265,000 deaths. Currently, Brazil is one of the most affected countries by COVID-19, with insufficient control measures implementation. Controlling the COVID-19 outbreak in Brazil is a challenge of continental proportions due to the population's size and urban density, inefficient maintenance of social distancing and testing strategies, and limited availability of testing resources [9].

This study addresses the COVID-19 testing prioritization for symptomatic patients to assist the early COVID-19 detection in Brazil. Addressing this problem is relevant due to the need for prioritization guidelines to improve testing and control strategies' efficiency. Therefore, the main Research Question (RQ) is: can demographic characteristics and symptoms that do not require expensive exams effectively assist the test prioritization for early COVID-19 detection in Brazil? From the main RQ, four Secondary RQ (SRQ) are: (1) what demographic characteristics are relevant to conduct the test prioritization? (2) what symptoms are suitable to drive the test prioritization? (3) what is the most suitable classification model for test prioritization? (4) what are the impacts of the reduction of reported symptoms in the test prioritization?

Aim of the Study

The study relied on pre-processing a raw dataset with information on 55,676 patients, aiming to provide a classification model that effectively recommends or not the prioritization of symptomatic patients for COVID-19 testing (i.e., a binary classification problem). The implementation of classification models also relied on supervised learning and the algorithms Multilayer Perceptron (MLP), Gradient Boosting Machine (GBM), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Linear Regression (LR). The algorithms were trained and tested using pre-processed datasets composed of demographic characteristics and reported symptoms that do not require expensive exams [10]. Usage of such symptoms is a relevant strategy for COVID-19 test prioritization due to the majority of the Brazilian population's high poverty levels [11].

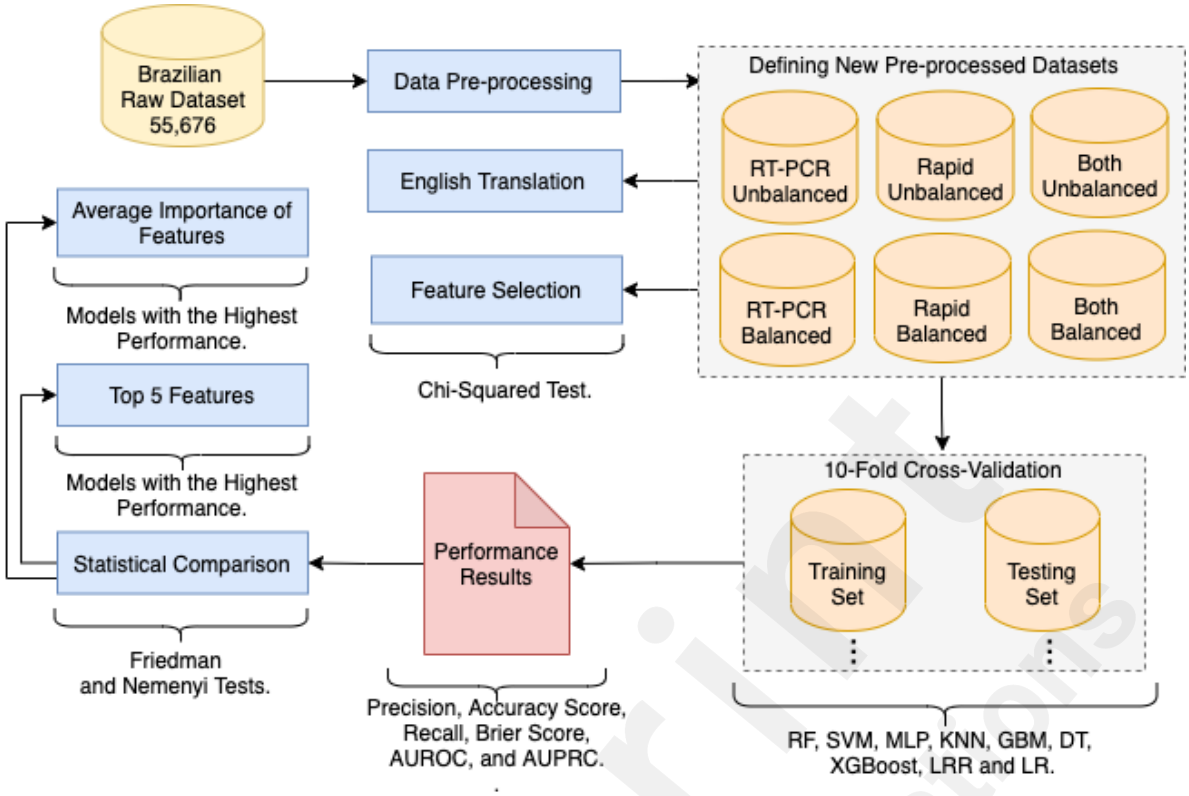
Our findings also provide insights for developers of e-Health and m-Health systems when choosing the most suitable classification model for COVID-19 testing prioritization. Such insights are also relevant for healthcare professionals and policymakers who envision applying a classification model to prioritize symptomatic patients for testing. The study enhances the state-of-the-art by providing three main contributions: (1) the pre-processing of raw data from 55,676 Brazilians, with the availability of data related to symptomatic patients [10]; (2) the implementation of classification models, along with reports of feature ranking, to support the COVID-19 test prioritization [12]; and (3) a comparative analysis of the classification models.

Methods

This study's research methodology consists of data pre-processing, the definition of new datasets, English translation, feature selection, 10-fold cross-validation, statistical comparisons, and feature ranking (Figure 1). The raw data from 55,676 Brazilians were pre-processed to define new datasets with information about symptomatic patients tested for COVID-19 using Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) and rapid tests (antibody and antigen). The textual descriptions of six pre-processed datasets (i.e., *RT-PCR Unbalanced*, *RT-PCR Balanced*, *Rapid Unbalanced*, *Rapid Balanced*, *Both Unbalanced*, and *Both Balanced*) were translated from Portuguese into English for public data availability. The Chi-squared test was applied in the new datasets to support the feature selection with a $P < .01$, verifying the relevance of features for the classification task by dependence and independence relations [13]. The chi-squared test for independence compared two variables in a contingency table to verify if they relate to each other.

We applied the 10-fold cross-validation method, with five repetitions, to validate the MLP, GBM, DT, RF, XGBoost, KNN, SVM, and LR (weak/strong regularization) classification models using the six datasets. We selected such algorithms because they have different characteristics, such as using neural layers, tree combinations, and calculating the distance between data. The mean results for classification metrics were also calculated: precision, accuracy score, recall, Brier score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). The recall results were further analyzed using the Friedman and Nemenyi statistical tests to improve the classification models' comparisons. We used the Friedman test to verify the differences between classification models. We applied the Nemenyi test to group classification models based on the verification of differences using multiple comparisons. Finally, we conducted features' ranking for each classification model with the highest performance using the permutation feature importance method, providing average importance and Standard Deviation (SD). The source code for replication is available in a GitHub repository [12].

Figure 1. Overview of the research methodology applied for the study. The methodological steps consist of data pre-processing, the definition of new datasets, English translation, feature selection, 10-fold cross-validation, statistical comparisons, and feature ranking.



Data Collection

The raw data from 55,676 Brazilians include information on tested patients in a spreadsheet format. However, the data collection is not a contribution of this study. The raw data was collected by the public health agency of the city of Campina Grande, Paraíba state, in Northeast Brazil. Such a public agency is informed by all the COVID-19 exams performed in the city of Campina Grande. The health agency employees removed patient identification, and the data made available were reused to enable this study. The raw dataset comprises categorical features such as *Health Professional*, *Security Professional*, *Ethnicity*, *Test Type*, *Fever*, *Sore Throat*, *Dyspnea*, *Olfactory Disorders*, *Cough*, *Coryza*, *Taste Disorders*, *Headache*, *Additional Symptoms*, *Test Result*, *Comorbidities*, *Test Status*, and *Symptoms Description*.

Data Pre-processing

We conducted the data pre-processing using the Python programming language. The raw dataset was pre-processed by applying string matching algorithms to correct inconsistencies. One example of inconsistency was the occurrence of empty columns of symptoms; however, the same symptoms were in a column for the general description of symptoms.

Besides, the following instances were removed due to our exclusion criteria: patients with uncompleted tests or undefined final classifications (12,929/55,676, 23.22%); duplicated instances (251/55,676, 0.45%); outliers related to input errors (10,408/55,676, 18.69%); test types that are not RT-PCR or rapid (771/55,676, 1.38%); undefined gender (27/55,676, 0.05%) and asymptomatic patients (11,269/55,676, 20.24%). Asymptomatic patients were removed because the inputs for the algorithms rely on demographic characteristics and symptoms.

Removing the feature related to the symptoms' description provides dimensionality reduction in the raw dataset feature space. For example, fatigue was removed because the symptom was reported by 228 (0.41%) of the 55,676 patients. Given the main focus on symptoms, the datasets did not include comorbidities and the remaining features (e.g., *Ethnicity*). As inclusion criteria, the most frequently reported symptoms (i.e., fever, sore throat, dyspnea, olfactory disorders, cough, coryza, taste

disorders, and headache) and relevant demographic characteristics (i.e., gender and health professional) were selected as features of unbalanced and balanced datasets (Table 1). Healthcare professionals were considered relevant due to the frequency of exposure to SARS-Cov-2. However, for gender, there is no consensus if there is a difference in the proportions of males and females infected with SARS-Cov-2 (usually a relatively even distribution) [14, 15].

The categorical data were converted into binary representation during the pre-processing. For the feature *Gender*, the number 0 represents a female patient, and 1 represents a male. For the features *Health Professional*, *Fever*, *Sore Throat*, *Dyspnea*, *Olfactory Disorders*, *Cough*, *Coryza*, *Taste Disorders*, and *Headache*, the number 0 represents positive response, and 1 represents negative response. For each dataset, the *Test Result* is the class that can be labeled as 0 for recommending a symptomatic patient for COVID-19 test prioritization or 1 for not recommending such patient's prioritization.

Table 1. Demographic and symptoms from symptomatic patients of both test types datasets.

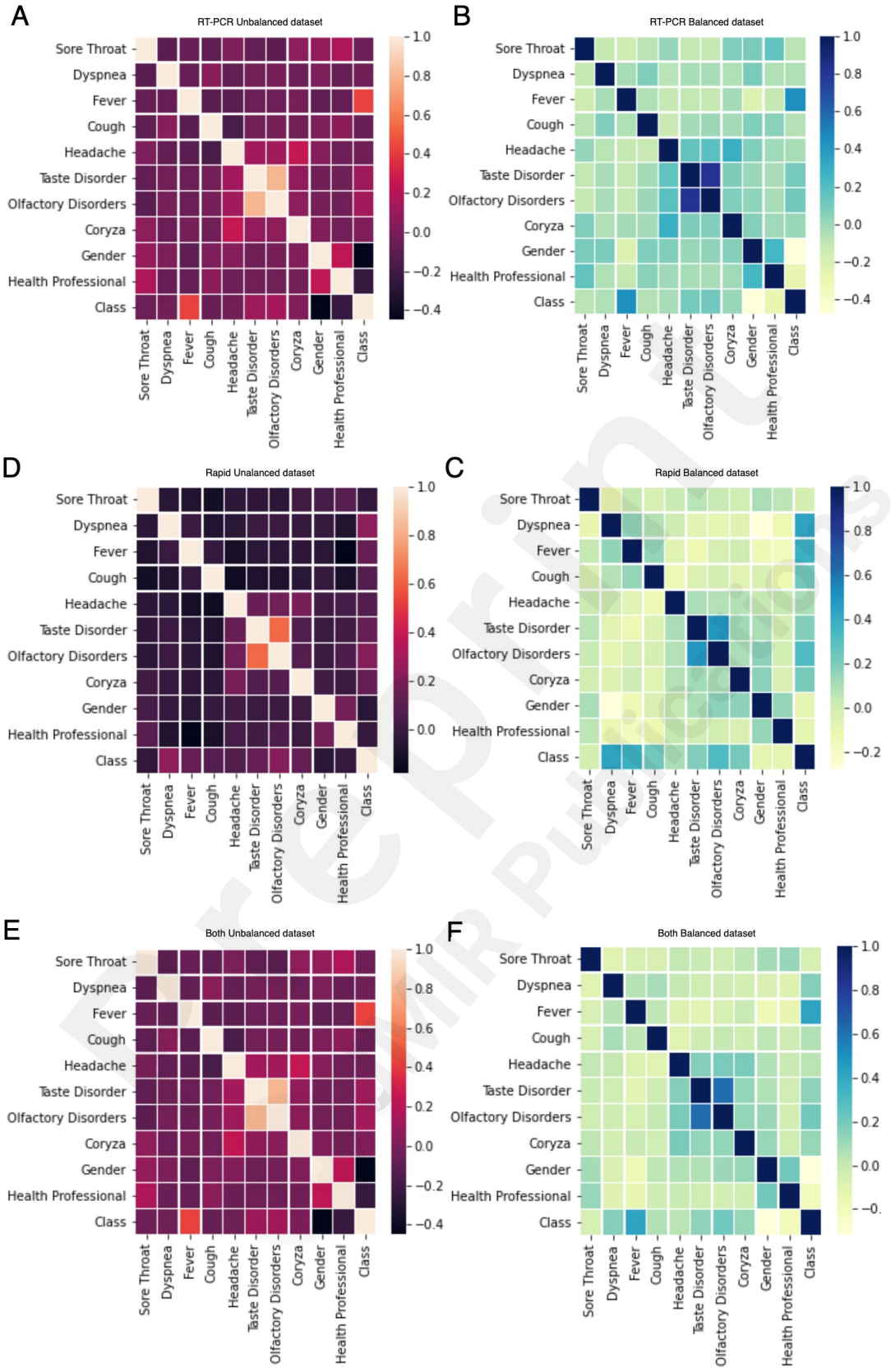
Features	Unbalanced (n=20,021)	Balanced (n=3,128)
Demographic characteristics		
Gender, males (%)	8,919 (44.55%)	1,639 (52.40%)
Health Professional, n (%)	2,485 (12.41%)	475 (15.19%)
Symptoms		
Fever, n (%)	9,169 (45.80%)	1,856 (59.34%)
Sore Throat, n (%)	5,976 (29.85%)	848 (27.11%)
Dyspnea, n (%)	3,704 (18.50%)	1,082 (34.59%)
Olfactory Disorders, n (%)	1,967 (9.82%)	522 (16.69%)
Cough, n (%)	11,641 (58.14%)	1,944 (62.15%)
Coryza, n (%)	1,159 (5.79%)	266 (8.50%)
Taste Disorders, n (%)	1,596 (12.37%)	387 (12.37%)
Headache, n (%)	4,034 (20.15%)	577 (18.45%)

The pre-processing included undersampling using the Near-miss technique [16], considering COVID-19 positive and negative cases. Undersampling was applied instead of oversampling to prevent the usage of synthetic data in training and testing sets. However, as stated, unbalanced data are also considered, without undersampling, to improve the experiments' representativity and achieve a scenario closer to a real-world setting, with more negative than positive COVID-19 cases.

Using the Chi-squared test for the *Both Unbalanced* and *Both Balanced* datasets, the independence hypothesis was only confirmed for *Headache*. For the *RT-PCR Unbalanced* dataset, the independence hypothesis was confirmed for *Sore Throat*, *Dyspnea*, *Headache*, and *Coryza*. In the *Rapid Unbalanced* dataset, the independence hypothesis was confirmed for *Sore Throat* and *Health Professional's* features. For the *RT-PCR Balanced* dataset, the independence hypothesis was confirmed for *Dyspnea*, *Cough*, *Headache*, and *Coryza*; while for the *Rapid Balanced*, the hypothesis was only confirmed for *Sore Throat*. Such information was used for feature selection

during the experiments, presenting scenarios with different numbers of symptoms to implement classification models. Besides, we used a correlation matrix to analyze the correlation coefficients between the features for each dataset (Figure 2). For example, *Fever* is among the features with the highest correlation coefficients for all datasets.

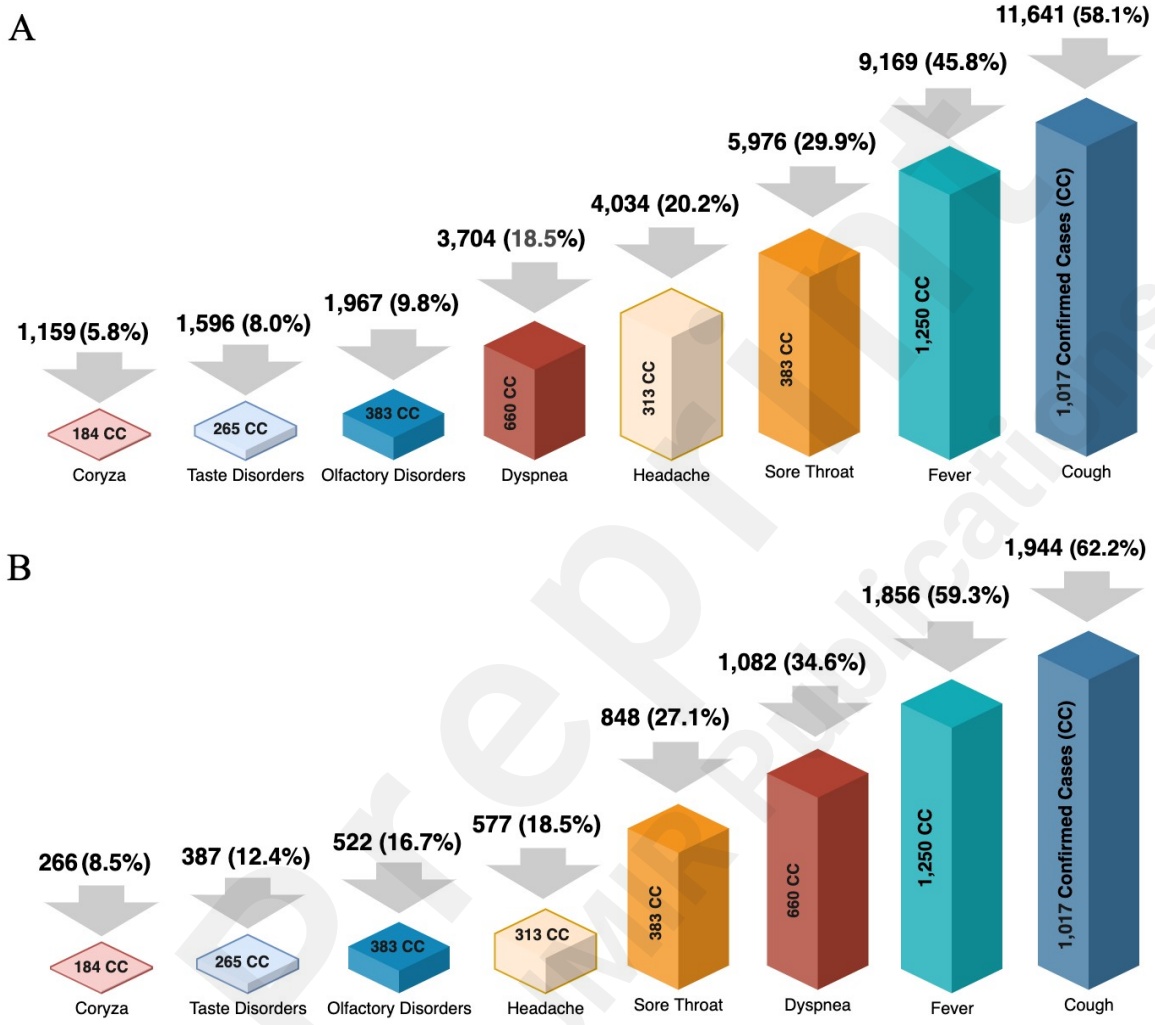
Figure 2. Correlation matrix for (A) *RT-PCR Unbalanced* dataset, (B) *RT-PCR Balanced* dataset, (C) *Rapid Unbalanced* dataset, (D) *Rapid Balanced* dataset, (E) *Both Unbalanced* dataset, and (F) *Both Balanced* dataset.



On the one hand, the *Both Unbalanced* dataset is composed of 20,021 patients tested by both RT-PCR and rapid tests. The reduction in the number of patients occurred due to the uncompleted tests, duplicated instances, outliers related to input errors, test type, and asymptomatic patients. The *Both Unbalanced* dataset contains 1.564 (7.81%) positive and 18,457 (92.19%) negative COVID-19 cases, while the balanced one included 1.564 cases of each class. From the female patients, 496 (2.48%) are positive and 10,606 (52.97%) negative cases. For male patients, 1,068 (5.33%) are positive and

7,851 (39.21%) negative cases. Cough was the most frequent symptom (11,641/20,021, 58.1%). The fever was the second most common symptom (9,169/20,021, 45.8%). The remaining symptoms were reported by at most 5,976 (29.9%) of the symptomatic patients (Figure 3A).

Figure 3. (A) The frequency of symptoms for the 20,021 symptomatic patients of the *Both Unbalanced* dataset and the number of Confirmed Cases (CC). Top values are frequencies; numbers on the geometric forms are the CC for frequency. (B) The frequency of symptoms for the 3,128 symptomatic patients of the *Both Balanced* dataset and the number of CC.



On the other hand, The *Both Balanced* dataset contains 3,128 patients tested by RT-PCR and rapid tests. The Near-miss technique reduced the number of negative cases to be equal to positive cases. 496 (15.86%) are positive and 993 (31.75%) negative cases from the female patients. For males, 1,068 (34.14%) are positive and 571 (18.25%) negative cases. Cough and fever continued to be the first and second most frequently reported symptoms, respectively. The remaining symptoms were also reported by at most 1,082 (34.6%) patients (Figure 3B).

Finally, the *RT-PCR Unbalanced* dataset included 916 (32.96%) positive and 1,863 (67.04%) negative COVID-19 cases, while the balanced one included 916 cases of each class. The *Rapid Unbalanced* dataset included 648 (3.76%) positive and 16,594 (96.24%) negative COVID-19 cases, while the balanced one included 648 of each class. The six scenarios' presentation aims to compare the classification models' results using various test types. Thus, there is no requirement to implement different clinical protocols or select patients with specific profiles for testing based on the results related to the six scenarios presented in this article.

Algorithms

We implemented the classification models using supervised learning and the MLP, GBM, DT, RF, XGBoost, KNN, SVM, and LR algorithms. An MLP Machine Learning (ML) algorithm [17] of one hidden layer learns the function

$$f(x)=W_2 g(W_1^T x+b_1)+b_2,$$

where W_1 represents the weights of the input layer, W_2 represents the hidden layer, b_1 is the bias added to the hidden layer, b_2 is the output layer, and g is the activation function.

The GBM is a fixed size decision tree that uses a boosting strategy [18]. This ML algorithm has a built-in feature selection and aims to provide the estimation or approximation $\hat{F}$ or the function $F*(x)$ that maps x to y, minimizing the expected value using a loss function $L(y,F(x))$ over the joint distribution [19], given by

$$F*¿ arg\,min_F E_{y,x} L(y,F(x))=arg\,min_F E_x[E_y(y,F(x))\vee x].$$

A DT is an ML algorithm that usually uses a divide and conquer strategy to generate a directed acyclic graph by applying division rules based on information gain [20]. The algorithm has a built-in feature selection, and the information gain is guided by the concept of entropy H, which measures the randomness of a discrete random variable A (with domain $a_1,a_2,\ldots,a_n$), given by

$$H(A)=-\sum_{i=1}^{n} p_i \log_2(p_i),$$

where p_i is the probability of observing each value $a_1,a_2,\ldots,a_n$. This algorithm enables a straightforward interpretation of results by following the decision rules of a unique tree.

The RF is an ML algorithm that relies on classification and regression trees, following specific tree growing rules, tree combination, self-testing, and post-processing [21]. The algorithm has a built-in feature selection, assessed by the Gini impurity criterion index. The binary split of a node n is given by

$$Gini(n)=1-\sum_{j=1}^{2} (p¿¿j)^2,¿$$

where p_j is the relative frequency of class j. This algorithm also enables a straightforward interpretation of results by following the decision rules of the trees.

As a variant of the GBM, the XGBoost is a regression tree with the same decision rules as a decision tree [22]. If the XGBoost ML algorithm consists of K decision trees, the optimization objective function is given by

$$\hat{y}_i=\sum_{k=1}^{K} f_k(x_i), f_k \in F,$$

where f_k is an independent tree with leaf scores, and F is the space of a regression tree. Both

algorithms enable a straightforward interpretation of results.

The KNN is a distance-based ML algorithm that identifies a new instance based on neighbors' distance [23]. An instance represents a point in the space, and the algorithm calculates the distance between two points using a metric such as the Euclidean distance, given by

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_i^l - x_j^l)^2},$$

where x_i and x_j are vectors representing objects in the space, and x_i^l and x_j^l the *l-th* elements of the vectors.

The SVM is an ML algorithm that handles binary data using a line to achieve the maximum distance between the data. The algorithm comprises four basic concepts: separation hyperplane, maximum margin hyperplane, soft margin, and kernel function [17]. For instance, the maximization of the margin hyperplane is given by

$$f(\vec{x}) = sqn¿$$

where y_i are the output variables, x_i are input vectors, *b* is the bias, *K* is a dot-products function (Kernel), and α_i is calculated by the maximization of

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j . K(\vec{x}_i, \vec{x}_j),$$

where x_j are the named support vectors when α_i is greater than 0.

Finally, the LR is an extension of linear regression that estimates relations between variables, using a sigmoid function during probabilistic classifications [24], given by

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

where *z* is the weighted sum of the evidence of a class. Regularization can also be used to prevent overfitting. We applied the LR algorithm to compare a compact and linear model's performance with the previous ML approaches.

We used the Python programming language and the SciPy library [25] to implement and validate the classification models based on such algorithms. We applied the random search method to configure the algorithms' hyperparameters to improve performance carefully. The configurations can be verified in the GitHub repository [12].

Classification Metrics

We calculated the precision, accuracy score, recall, Brier score, AUROC, and AUPRC for the classification models [26]. The precision represents the proportion of classifications that are true positives and is given by

$$Precision = \frac{TP}{TP + FP},$$

where *TP* is the true positives and *FP* is the false positives. The accuracy score presents fractions of correct classifications and is given by

$$A(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i),$$

where *A* is the accuracy score, $\hat{y}_i$ is the classified value of a sample, y_i is the corresponding true value, *n* is the number of samples, and $I(x)$ is the indicator function.

The recall calculates the actual positives correctly positives and is given by

$$Recall = \frac{TP}{TP + FN},$$

where *FN* is the number of false negatives. It is relevant for evaluating classifications related to diagnosis due to the highly undesired impacts of false negatives.

The Brier score provides the mean squared difference between predicted probabilities and expected results, given by

$$Brier = \frac{1}{n} \sum_{t=1}^{n} (f_t - o_t)^2,$$

where f_t is the predicted value, o_t is the expected value, and *n* is the number of samples.

Finally, the AUROC provides an overview of the diagnostic abilities of the models. However, the usage of the AUPRC is usually recommended when handling problems using unbalanced data.

Results

The implementations of classification models using the MLP, GBM, DT, RF, XGBoost, KNN, SVM, and LR algorithms are available in the GitHub repository [12]. Using the 10-fold cross-validation with five repetitions, the mean values of precision, accuracy score, recall, and Brier score of the decision-tree-based classification models are among the best results (Table 2). Such models presented similar results using the six datasets. For the *RT-PCR Unbalanced/Balanced* and *Both Unbalanced/Balanced* datasets, the LR algorithm was outperformed by the other models. In the results, LR and LRR stand for models with weak and strong regularization, respectively.

Table 2. Results of 10-fold cross-validation for the classification models using the unbalanced and balanced datasets.

Datasets and Models	Precision %	Accuracy Score %	Recall %	Brier Score
RT-PCR Unbalanced and Balanced				
MLP, unbalanced (balanced)	97.33 (95.86)	96.24 (95.81)	97.08 (95.80)	0.04 (0.04)
GBM, unbalanced (balanced)	97.32 (95.95)	96.30 (95.70)	97.17 (95.47)	0.04 (0.04)

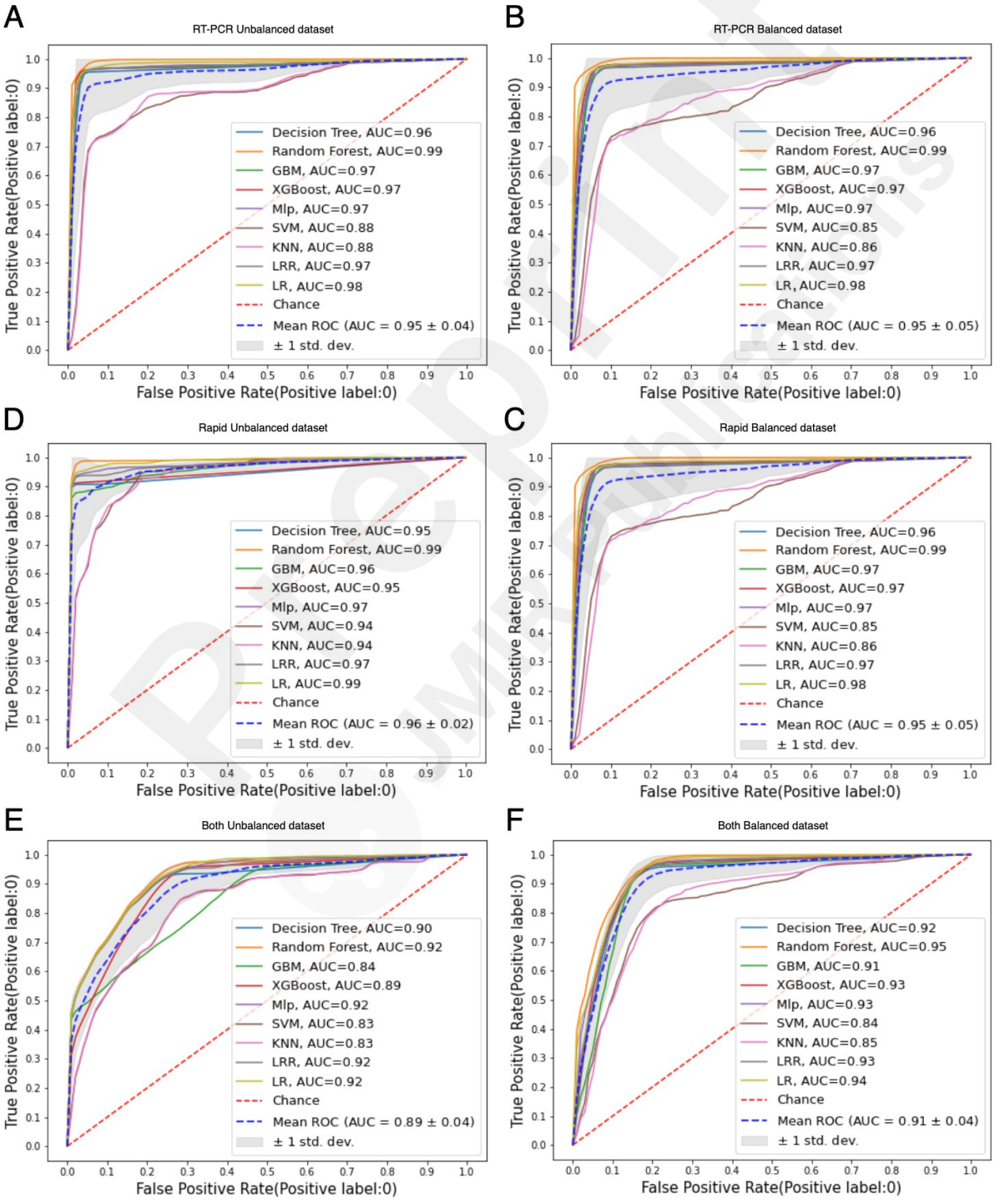
RF, unbalanced (balanced)	97.42 (96.06)	96.55 (96.00)	97.46 (95.97)	0.04 (0.04)
DT, unbalanced (balanced)	97.49 (96.50)	96.33 (95.91)	97.04 (95.32)	0.04 (0.04)
XGBoost, unbalanced (balanced)	97.36 (95.94)	96.30 (95.52)	97.13 (95.10)	0.04 (0.04)
KNN, unbalanced (balanced)	97.38 (95.92)	96.55 (95.48)	97.50 (95.04)	0.03 (0.05)
SVM, unbalanced (balanced)	97.17 (95.84)	96.19 (95.58)	97.18 (95.34)	0.04 (0.04)
LRR, unbalanced (balanced)	86.97 (76.86)	86.72 (81.70)	94.37 (90.93)	0.13 (0.18)
LR, unbalanced (balanced)	87.00 (76.56)	86.72 (80.63)	94.33 (88.53)	0.13 (0.19)
Rapid Unbalanced and Balanced				
MLP, unbalanced (balanced)	99.33 (96.66)	98.70 (95.40)	99.32 (94.10)	0.01 (0.05)
GBM, unbalanced (balanced)	99.33 (96.18)	98.72 (95.33)	99.34 (94.50)	0.01 (0.05)
RF, unbalanced (balanced)	99.26 (96.42)	98.76 (95.21)	99.44 (93.98)	0.01 (0.05)
DT, unbalanced (balanced)	99.37 (95.51)	98.69 (94.59)	99.27 (93.67)	0.01 (0.05)
XGBoost, unbalanced (balanced)	99.33 (96.83)	98.72 (95.41)	99.34 (93.94)	0.01 (0.05)
KNN, unbalanced (balanced)	99.31 (97.43)	98.84 (94.58)	99.49 (91.63)	0.01 (0.05)
SVM, unbalanced (balanced)	99.30 (97.30)	98.73 (95.60)	99.37 (93.85)	0.01 (0.04)
LRR, unbalanced (balanced)	96.65 (82.00)	96.23 (84.22)	99.53 (87.93)	0.04 (0.16)
LR, unbalanced (balanced)	96.75 (84.75)	96.14 (85.33)	99.32 (86.32)	0.04 (0.15)
Both Unbalanced and Balanced				
MLP, unbalanced (balanced)	95.36 (93.53)	94.82 (89.18)	99.20 (84.23)	0.05 (0.11)
GBM, unbalanced (balanced)	95.23 (93.67)	94.73 (89.31)	99.25 (84.38)	0.05 (0.11)
RF, unbalanced (balanced)	95.31 (93.81)	94.87 (89.22)	99.32 (84.04)	0.05 (0.11)
DT, unbalanced (balanced)	95.43 (93.75)	94.79 (89.12)	99.10 (83.87)	0.05 (0.11)
XGBoost, unbalanced (balanced)	95.32 (93.60)	94.78 (89.22)	99.21 (84.24)	0.05 (0.11)
KNN, unbalanced (balanced)	95.50 (92.77)	91.09 (88.63)	94.81 (83.86)	0.09 (0.11)
SVM, unbalanced (balanced)	95.21 (93.36)	94.75 (89.33)	99.30 (84.73)	0.05 (0.11)
LRR, unbalanced (balanced)	92.45 (80.79)	92.04 (80.48)	99.48 (80.11)	0.08 (0.20)
LR, unbalanced (balanced)	92.49 (82.44)	91.98 (81.08)	99.36 (79.14)	0.08 (0.19)

When removing features according to the Chi-squared results, there is a considerable decrease in classification models' performance (Multimedia Appendix 1). However, in general, the classification models continue presenting good performances. For example, the KNN classification model presented the lowest accuracy score (77.42%) using the *RT-PCR Balanced* dataset. The remaining classification models, considering all datasets, presented accuracy scores between 80.15% and 97.58%. Depending on the pre-processed dataset, the LR (weak/strong regularization) continued to be outperformed by the other algorithms. Presenting such scenarios is relevant to analyze how the

algorithms behave when models are implemented with reduced reported symptoms.

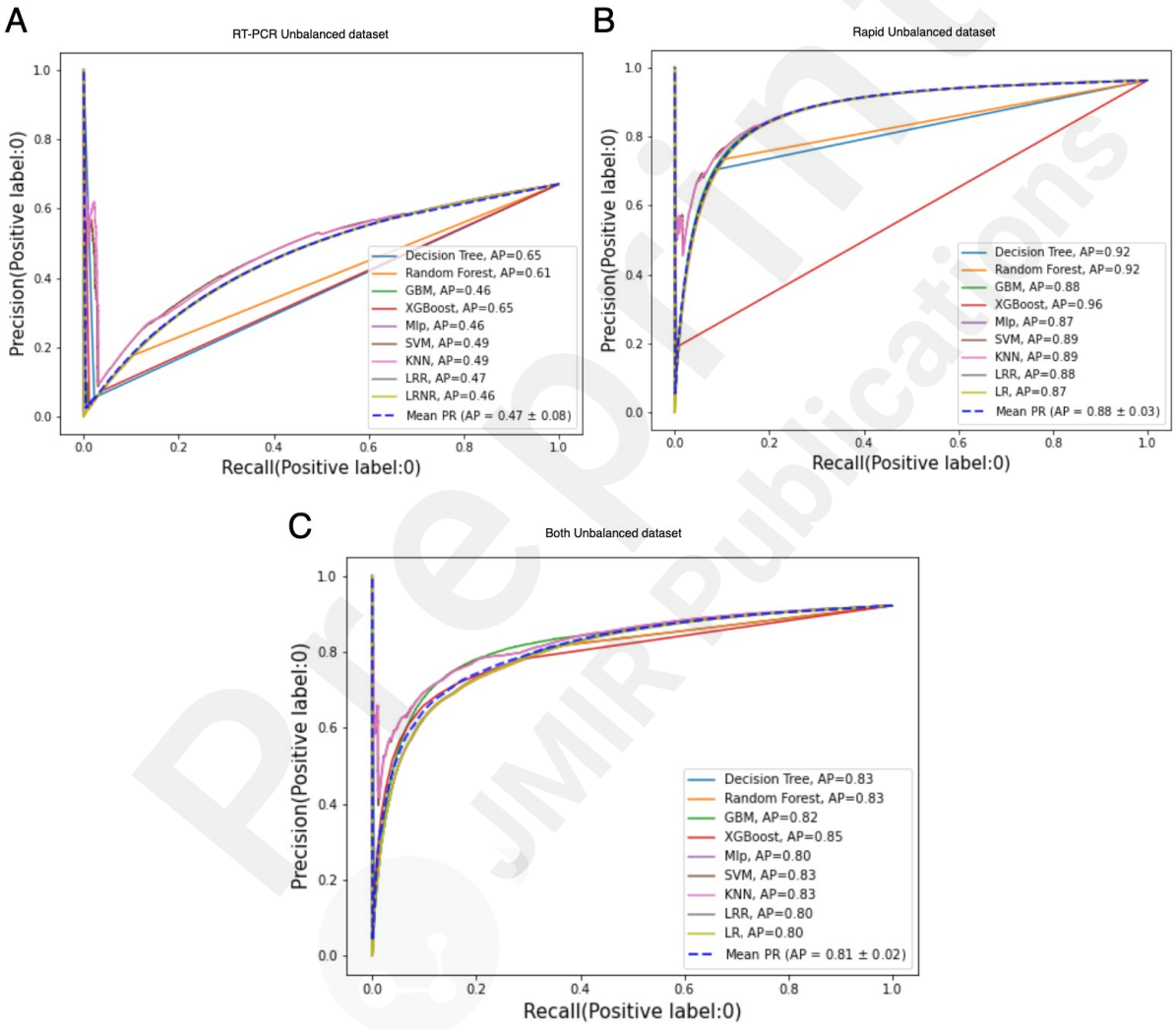
Also, by computing the AUROC using the RT-PCR, rapid, and both test scenarios, the trade-offs between sensitivity (true positive rate) and probability (false positive rate) were identified, evidencing the diagnostic abilities of the classification models when the discrimination threshold is varied (Figure 4). The classification models presented high discriminatory power for all scenarios, with the curves closer to each graphic representation's upper left corner. However, for such scenarios, the KNN and SVM classification models presented the lowest discriminatory power.

Figure 4. Models' ROC curves with (A) *RT-PCR Unbalanced*, (B) *RT-PCR Balanced*, (C) *Rapid Unbalanced*, (D) *Rapid Balanced*, (E) *Both Unbalanced*, and (F) *Both Balanced*.



Given the three unbalanced datasets, there are more negative than positive COVID-19 cases. We computed the AUPRC to verify the classification models when handling the minority class, analyzing the trade-off between precision and recall for different decision thresholds (Figure 5). The AUPRC was summarized using the Average Precision (AP), as a weighted mean of precision. The *RT-PCR Unbalanced* dataset is mildly unbalanced, with a baseline AUPRC of 0.33. The *Rapid Unbalanced* dataset is highly unbalanced, with a baseline AUPRC of 0.04. This is also the case for the *Both Unbalanced* dataset, with a baseline AUPRC of 0.08. The DT and XGBoost achieved the best AP value (65%) using the *RT-PCR Unbalanced* dataset. For the remaining scenarios, the classification models presented AP values between 80% and 96%.

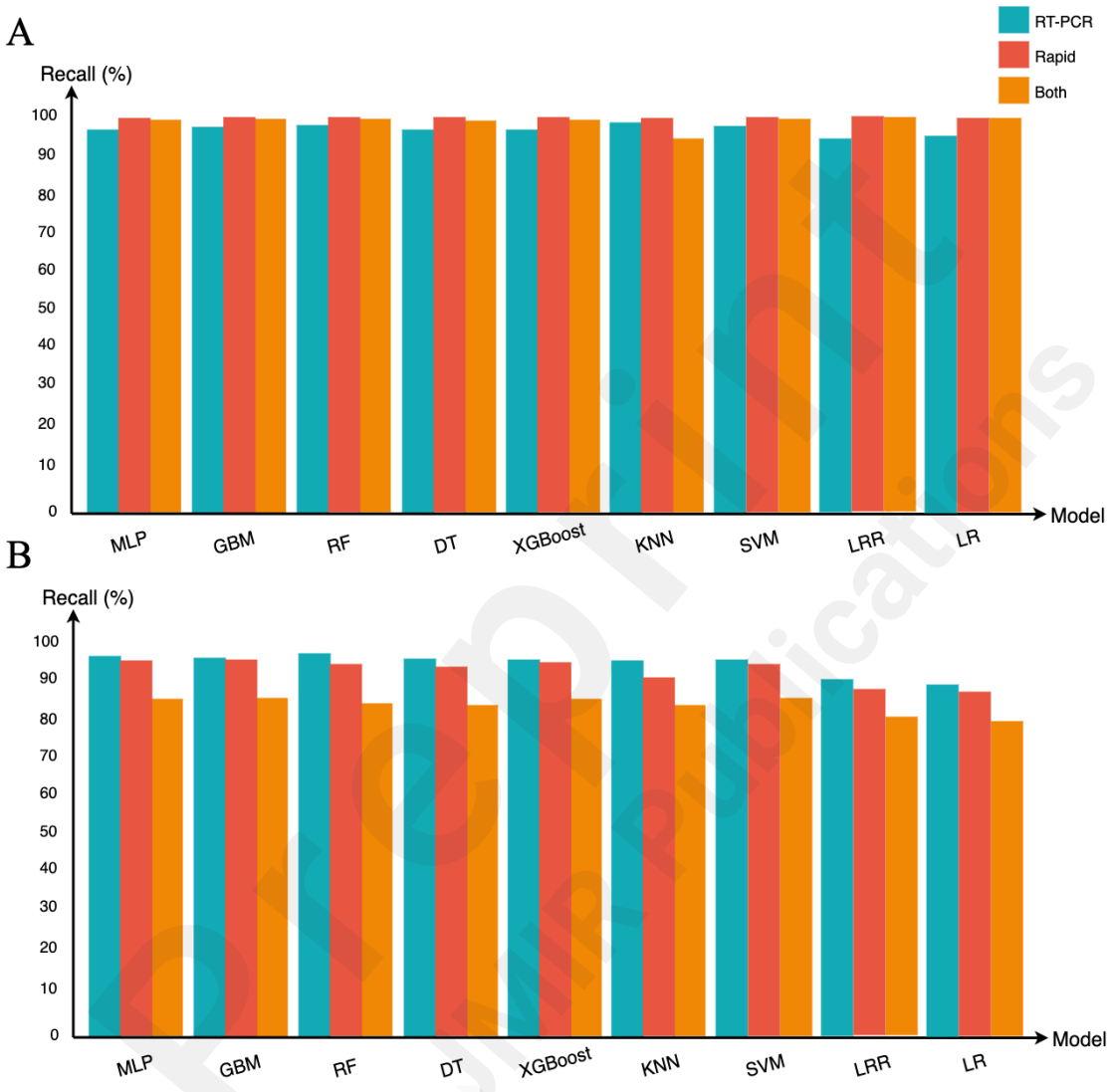
Figure 5. Models' PRC with (A) *RT-PCR Unbalanced* dataset, (B) *Rapid Unbalanced* dataset, and (C) *Both Unbalanced* dataset.



We also applied the Friedman and Nemenyi tests to improve confidence in evaluating the classification models, observing that the experiments' results are statistically significant. The classification models were compared over the six datasets using the Friedman test [27]. This comparison focused on the recall results due to the highly undesired impacts of false negatives in the COVID-19 application scenario (Figure 6). The null hypothesis is that all classification models are equivalent and have equal mean ranks. The tests resulted in a $P < .001$ for the *RT-PCR Unbalanced* ($t = 307.16$), *RT-PCR Balanced* ($t = 328.72$), *Rapid Unbalanced* ($t = 247.43$), *Rapid Balanced* ($t = 239.20$), *Both Unbalanced* ($t = 226.98$), and *Both Balanced* ($t = 343.10$). The results showed that the difference between the mean recall values is probably real ($P \leq .1$). The Friedman test ranked the

classification models for each dataset, resulting in an average rank for each classification model.

Figure 6. (A) The mean recall for the MLP, GBM, RF, DT, XGBoost, KNN, SVM, LRR, and LR classification models using the unbalanced datasets for RT-PCR, rapid, and both types. (B) The mean recall for the MLP, GBM, RF, DT, XGBoost, KNN, SVM, LRR, and LR classification models using the balanced datasets for RT-PCR, rapid, and both types.



Based on the Friedman test results, the Nemenyi test [27] was applied to compare the classification models using the mean ranks. The Critical Difference (CD) between the classification models was verified using the Nemenyi test with $\alpha = .1$. The CD is relevant to highlight if the classification models are separated by an interval less than the CD, meaning that the classification models are statistically indistinguishable. Thus, for most of the datasets, the difference between LRR/LR (statistically indistinguishable) and the other classification models is highlighted by the CD using the mean recall results (Multimedia Appendix 2). Depending on the dataset, MLP and GBM are also statistically indistinguishable, as is the case of DT, RF, XGBoost, KNN, and SVM.

From the classification metrics results and the Friedman and Nemenyi tests (Figure 6), the top five features of the classification models with the highest performances (i.e., MLP, GBM, DT, RF, XGBoost, and SVM) were ranked using the permutation feature importance method. Each average importance and SD values are presented for the decision-tree-based classification models and the RT-PCR, rapid, and both types scenarios (Table 3). The average importance and SD information relate to

reducing the feature importance when a feature is not considered. For example, according to the frequency of symptoms and the number of confirmed cases (Figure 3), *fever* showed higher average importance values for almost all scenarios than other reported symptoms. We also applied the permutation feature importance method for the unbalanced datasets (Multimedia Appendix 3).

Table 3. The average importance and SD values for each feature for the decision-tree-based classification models using the balanced datasets.

Datasets and Features	GBM	DT	RF	XGBoost
RT-PCR Balanced				
Gender, mean (SD)	0.233 (0.013)	0.245 (0.013)	0.238 (0.013)	0.233 (0.013)
Health Professional, mean (SD)	0.045 (0.005)	0.055 (0.005)	0.051 (0.005)	0.042 (0.005)
Fever, mean (SD)	0.260 (0.012)	0.263 (0.012)	0.267 (0.013)	0.262 (0.013)
Sore Throat, mean (SD)	0.100 (0.007)	0.102 (0.007)	0.100 (0.007)	0.101 (0.006)
Dyspnea, mean (SD)	0.096 (0.007)	0.100 (0.007)	0.098 (0.007)	0.092 (0.007)
Olfactory Disorders, mean (SD)	0.026 (0.004)	0.029 (0.003)	0.014 (0.004)	0.016 (0.003)
Cough, mean (SD)	0.087 (0.007)	0.096 (0.007)	0.084 (0.007)	0.082 (0.008)
Coryza, mean (SD)	0.026 (0.004)	0.027 (0.004)	0.022 (0.004)	0.014 (0.003)
Taste Disorders, mean (SD)	0.030 (0.004)	0.040 (0.004)	0.027 (0.004)	0.024 (0.003)
Headache, mean (SD)	0.021 (0.004)	0.024 (0.005)	0.018 (0.004)	0.002 (0.003)
Rapid Balanced				
Gender, mean (SD)	0.135 (0.009)	0.109 (0.009)	0.122 (0.010)	0.123 (0.010)
Health Professional, mean (SD)	0.026 (0.005)	0.027 (0.005)	0.020 (0.004)	0.027 (0.005)
Fever, mean (SD)	0.120 (0.012)	0.124 (0.012)	0.114 (0.011)	0.109 (0.010)
Sore Throat, mean (SD)	0.019 (0.005)	0.030 (0.005)	0.023 (0.004)	0.013 (0.004)
Dyspnea, mean (SD)	0.184 (0.012)	0.179 (0.012)	0.187 (0.013)	0.179 (0.013)
Olfactory Disorders, mean (SD)	0.175 (0.012)	0.178 (0.013)	0.180 (0.014)	0.154 (0.012)
Cough, mean (SD)	0.076 (0.009)	0.084 (0.011)	0.080 (0.008)	0.080 (0.008)
Coryza, mean (SD)	0.090 (0.008)	0.092 (0.009)	0.087 (0.007)	0.052 (0.005)
Taste Disorders, mean (SD)	0.078 (0.008)	0.081 (0.009)	0.053 (0.007)	0.071 (0.007)
Headache, mean (SD)	0.035 (0.007)	0.030 (0.007)	0.035 (0.006)	0.035 (0.007)
Both Balanced				
Gender, mean (SD)	0.159 (0.007)	0.160 (0.007)	0.153 (0.007)	0.156 (0.007)
Health Professional, mean (SD)	0.025 (0.004)	0.024 (0.004)	0.023 (0.004)	0.025 (0.004)
Fever, mean (SD)	0.211 (0.010)	0.215 (0.011)	0.213 (0.010)	0.209 (0.010)
Sore Throat, mean (SD)	0.080 (0.005)	0.081 (0.005)	0.082 (0.005)	0.078 (0.005)
Dyspnea, mean (SD)	0.077 (0.006)	0.075 (0.005)	0.073 (0.005)	0.076 (0.005)
Olfactory Disorders, mean (SD)	0.059 (0.006)	0.050 (0.005)	0.050 (0.005)	0.046 (0.005)
Cough, mean (SD)	0.060 (0.005)	0.058 (0.005)	0.054 (0.005)	0.060 (0.005)
Coryza, mean (SD)	0.047 (0.004)	0.042 (0.003)	0.040 (0.003)	0.044 (0.004)
Taste Disorders, mean (SD)	0.063 (0.006)	0.079 (0.006)	0.072 (0.005)	0.069 (0.005)
Headache, mean (SD)	0.042 (0.005)	0.046 (0.005)	0.044 (0.005)	0.045 (0.005)

We also present the results achieved using the permutation feature importance method for detailing the feature ranking for classifications with MLP and SVM models (Table 4). For example, similar to the decision-tree-based classification models', *Fever* presented higher average importance values for almost all test scenarios than other symptoms reported by patients. For such algorithms, we also present the average importance and SD for the unbalanced datasets (Multimedia Appendix 3).

Table 4. The average importance and SD for each feature for the MLP and SVM models and the balanced datasets.

Datasets and Features	MLP	SVM
RT-PCR Balanced		

Gender, mean (SD)	0.236 (0.013)	0.230 (0.013)
Health Professional, mean (SD)	0.048 (0.005)	0.042 (0.005)
Fever, mean (SD)	0.262 (0.013)	0.257 (0.013)
Sore Throat, mean (SD)	0.103 (0.006)	0.098 (0.006)
Dyspnea, mean (SD)	0.096 (0.007)	0.088 (0.007)
Olfactory Disorders, mean (SD)	0.027 (0.004)	0.015 (0.004)
Cough, mean (SD)	0.084 (0.008)	0.078 (0.007)
Coryza, mean (SD)	0.025 (0.004)	0.013 (0.004)
Taste Disorders, mean (SD)	0.031 (0.004)	0.020 (0.004)
Headache, mean (SD)	0.023 (0.003)	0.002 (0.003)
Rapid Balanced		
Gender, mean (SD)	0.120 (0.009)	0.117 (0.010)
Health Professional, mean (SD)	0.033 (0.006)	0.029 (0.005)
Fever, mean (SD)	0.115 (0.010)	0.105 (0.011)
Sore Throat, mean (SD)	0.012 (0.005)	0.023 (0.004)
Dyspnea, mean (SD)	0.177 (0.013)	0.177 (0.014)
Olfactory Disorders, mean (SD)	0.157 (0.012)	0.149 (0.012)
Cough, mean (SD)	0.082 (0.009)	0.076 (0.008)
Coryza, mean (SD)	0.064 (0.006)	0.058 (0.006)
Taste Disorders, mean (SD)	0.072 (0.007)	0.055 (0.006)
Headache, mean (SD)	0.036 (0.006)	0.028 (0.005)
Both Balanced		
Gender, mean (SD)	0.161 (0.007)	0.154 (0.007)
Health Professional, mean (SD)	0.025 (0.004)	0.024 (0.004)
Fever, mean (SD)	0.207 (0.010)	0.193 (0.009)
Sore Throat, mean (SD)	0.084 (0.005)	0.075 (0.005)
Dyspnea, mean (SD)	0.078 (0.006)	0.088 (0.006)
Olfactory Disorders, mean (SD)	0.055 (0.006)	0.049 (0.005)
Cough, mean (SD)	0.062 (0.005)	0.071 (0.005)
Coryza, mean (SD)	0.035 (0.003)	0.046 (0.003)
Taste Disorders, mean (SD)	0.071 (0.006)	0.068 (0.005)
Headache, mean (SD)	0.051 (0.005)	0.045 (0.005)

Therefore, the top five most significant features vary depending on the algorithm used to implement the classification model (Table 5). For the *RT-PCR Balanced* dataset, all algorithms prioritized the same top two features (i.e., *Fever* and *Gender*), slightly differing in the top three and top five; while, for the *Rapid Balanced* dataset, all algorithms prioritized the same top two features (i.e., *Dyspnea* and *Olfactory disorders*), also slightly different in the top three, top four, and top five. For the *Both Balanced* dataset, the algorithms prioritized the top two features similar to the classifications with the *RT-PCR Balanced* dataset. We also applied the permutation feature importance method to rank features using the unbalanced datasets (Multimedia Appendix 3).

Table 5. The five most significant for COVID-19 test prioritization using the classification models with the highest performances and the datasets.

Datasets and Models	Top One	Top Two	Top Three	Top Four	Top Five
RT-PCR Balanced					
MLP	Fever	Gender	Sore Throat	Dyspnea	Cough
GBM	Fever	Gender	Sore Throat	Dyspnea	Cough
RF	Fever	Gender	Sore Throat	Dyspnea	Cough
DT	Fever	Gender	Sore Throat	Dyspnea	Cough
XGBoost	Fever	Gender	Sore Throat	Dyspnea	Cough

SVM	Fever	Gender	Sore Throat	Dyspnea	Cough
Rapid Balanced					
MLP	Dyspnea	Olfactory Disorders	Gender	Fever	Cough
GBM	Dyspnea	Olfactory Disorders	Gender	Fever	Coryza
RF	Dyspnea	Olfactory Disorders	Gender	Fever	Coryza
DT	Dyspnea	Olfactory Disorders	Fever	Gender	Coryza
XGBoost	Dyspnea	Olfactory Disorders	Gender	Fever	Cough
SVM	Dyspnea	Olfactory Disorders	Gender	Fever	Cough
Both Balanced					
MLP	Fever	Gender	Sore Throat	Dyspnea	Taste Disorders
GBM	Fever	Gender	Sore Throat	Dyspnea	Taste Disorders
RF	Fever	Gender	Sore Throat	Dyspnea	Taste Disorders
DT	Fever	Gender	Sore Throat	Taste Disorders	Dyspnea
XGBoost	Fever	Gender	Sore Throat	Dyspnea	Taste Disorders
SVM	Fever	Gender	Dyspnea	Sore Throat	Cough

Also, to improve the experiments conducted to assist the COVID-19 test prioritization, we combined the classification models to define voting ensemble models using the majority voting strategy (Multimedia Appendix 4). Two combinations of classification models were considered for each dataset: decision-tree-based models (i.e., GBM, DT, RF, and XGBoost) and non-decision tree models (i.e., MLP, SVM, KNN, LRR, and LR). In general, for the voting ensemble models implemented with the six datasets, the mean results of classification metrics using 10-fold cross-validation were similar to those of MLP, GBM, DT, RF, XGBoost, KNN, SVM, LRR, and LR models (Table 2).

Discussion

Principal Findings

The raw dataset's data pre-processing enabled the implementation, validation, and comparison of classification models with different characteristics, such as using neural layers, tree combinations, and calculating the distance between data. The pre-processing also resulted in the public data availability of symptomatic patients tested using RT-PCR and rapid tests [10]. Thus, the datasets can be reused by other studies to improve the state-of-the-art.

The algorithms were trained and tested using the unbalanced and balanced datasets, improving data representativity. The best classification metrics results were related to the RT-PCR and rapid tests scenarios using unbalanced and balanced data. Although the classification models' performance was similar for the RT-PCR and rapid tests scenarios, the RT-PCR test scenario is the most clinically relevant one due to the RT-PCR testing's high confidence. The RT-PCR test's precision increases confidence in the diagnosis, even if the patient was tested in the first days after symptoms onset. For both test scenarios with unbalanced data, although presenting a low Brier score and high precision, accuracy score, and recall, the classification models presented a lower AUROC because of the higher negative than positive COVID-19 cases. For both test scenarios with balanced data, the Brier score continued to be low. The precision, accuracy, and AUROC were higher; however, the recall results were slightly decreased if compared to the unbalanced data results.

The recall metric is relevant due to the undesired impacts of false negatives in clinical practice. Thus, we improved the classification models' quality of comparisons by applying the Friedman and Nemenyi tests based on the six datasets' recall. We used such statistical comparison results for defining the MLP, GBM, DT, RF, XGBoost, and SVM as the classification models with the highest performances for COVID-19 test prioritization in Brazil.

Given the classification models with the highest performances and the five most significant features for COVID-19 test prioritization, the fever's importance as one of the top two features is according to the statistics presented above (Figure 3). The statistics showed that fever was the second most frequent symptomatic patient reported, confirmed as COVID-19 cases. *Gender* and *Dyspnea* were also among the highest-ranked features used by classification models. For example, for the *RT-PCR Balanced* dataset, observing the DT model's decision rules to get an overview of the role of gender in classifications, positive or negative decisions for males and females differ based on reported symptoms and the *Health Professional* feature. However, further investigation about the role of gender in classifications is recommended for future works.

Therefore, SRQ 1 was answered by showing that *Gender* and *Health Professional* features are related to relevant demographic characteristics to support the COVID-19 test prioritization in Brazil (Table 4 and Table 5). The SRQ 2 was also answered, showing that fever, sore throat, dyspnea, olfactory disorders, cough, coryza, taste disorders, and headache are relevant symptoms.

All decision-tree-based classification models considered in this study are among the classification models with the highest performances, grouped based on the results of classification metrics and statistical tests. This fact is relevant due to the high levels of decision trees' interpretability, positively impacting healthcare professionals' final decision-making. In clinical practice, ML-based applications' acceptance increases when healthcare professionals can easily understand/interpret classification models' outputs to track decision-making logic [29]. Given the grouping of models with similar performances, we used the criterion of easy interpretability to answer the SRQ 3. Thus, the DT classification model was considered the most suitable for the COVID-19 test prioritization in Brazil. We configured the model with: the Gini impurity criterion, best split strategy, no maximum depth, a minimum number of two samples split and one sample leaf, no minimum weighted fraction leaves and no impurity decrease and split, unlimited number of features and leaves, global random state instance, no class weight, and no pruning. As one of the classification models with the highest performances, DT provides a simple tree representation of the decision-making, enabling a unique tree's straightforward interpretation by healthcare professionals.

To answer the SRQ 4, we analyzed the DT model's classification results, observing that a

considerable fraction of the incorrectly classified instances occurred when patients reported only one, two, or three symptoms. Besides, we conducted an experiment to verify the impacts of reducing features in the performance of the implemented classification models (Multimedia Appendix 1). For example, with the *Both RT-PCR Balanced* dataset, when the symptoms of sore throat, dyspnea, headache, and coryza are not considered to implement the DT classification model, the performance results are decreased considerably. This reduces the ability of the model to distinguish between positive and negative cases.

Although the DT is considered the most suitable model, all the other classification models that presented high performance are relevant to address the COVID-19 test prioritization. In Brazil, due to other epidemics (e.g., dengue fever [30]), many people report symptoms that may or may not be related to COVID-19. As a developing country, Brazil also suffers from inefficient testing strategies, such as shortages of COVID-19 tests. One of the available classification models can be applied for COVID-19 test prioritization during primary healthcare, with a mean accuracy score of at least 88.63%.

Comparison With Prior Work

The relevance of researches addressing viral infection outbreaks is evidenced from the public administration (e.g., surveillance systems) to the diagnosis viewpoint. For example, Son et al. [31] used a South Korean time series of influenza incidence for early outbreak detection, aiming to assist the definition of control policies. Chatterjee, Gerdes, and Martinez [32] analyzed COVID-19 datasets to identify risks of spreading, identify correlated factors associated with the disease's spread, identify the impact of social isolation, and experiment with univariate long short term memory models for forecasting of total cases and total deaths. In general, infectious disease research is guided by trends in data analytics [33].

Indeed, the COVID-19 pandemic is an example of a problematic scenario. Kumar [34] applied cluster analysis to study and improve the monitoring of SARS-Cov-2 infections in India, providing insights on clusters of affected Indian states and union territories. Besides aiming to improve the management of available resources, Khakharia et al. [37] developed outbreak classification models for COVID-19 using datasets with information about patients who live in India, Bangladesh, the Democratic Republic of Congo, Pakistan, China, Philippines, Germany, Indonesia, Ethiopia, and Nigeria. Vaid et al. [38] implemented and validated models (e.g., XGBoost) to predict mortality and critical events using electronic health records of patients who tested positive for COVID-19 in New York City.

To assist the COVID-19 detection, Brinati et al. [37] validated models implemented using DT, Extremely Randomized Trees (ERT), KNN, LR, Naive Bayes (NB), RF, and Three-Way Random Forest (TWRF) algorithms. The authors considered COVID-19 detection using routine blood exams, gender, and age. The accuracy of the models ranged between 82% and 86%. However, the large number of required blood exams (i.e., 13) is a limitation, which may compromise this approach's feasibility in low and middle-income countries.

Ahamad et al. [21] used a Chinese dataset to assist the COVID-19 detection considering symptoms (i.e., fever, cough, pneumonia, lung infection, coryza, muscle soreness, and diarrhea), gender, age, travel history, and isolation. The authors validated the XGBoost, SVM, DT, RF, and GBM models. XGBoost presented the highest accuracy with more than 85%, varying according to age. However, lung infection usage, detected by chest images, increases costs and may limit the disease's rapid screening.

Aiming to improve confidence in screening COVID-19, Mei et al. [28] used Computerized Tomography (CT) images along with symptoms (i.e., fever, cough, and cough with sputum), exposure history, laboratory testing (i.e., white blood cells, neutrophils, percentage neutrophils, lymphocytes, and percentage lymphocytes), age, gender, and temperature. They applied the deep convolutional neural network to analyze images, besides comparing the performance of SVM, RF, and MLP models, showing that MLP presented the highest accuracy score. Afterward, the authors combined images and clinical information. Similarly, requiring images increases costs and may limit the rapid screening of COVID-19 in low- and middle-income countries.

Finally, Zoabi, Deri-Rozov, and Shomron [4] used gender, age, symptoms (e.g., cough, fever, sore throat, shortness of breath, and headache), and contact with a confirmed case, to classify positive and negative COVID-19 cases. The authors implemented a GBM model based on data reported by the Israeli Ministry of Health. The GBM model presented an AUROC of 86% and 90% using, respectively, a reduced set of features and the complete set. Similarly to our study, the authors reported the high importance of gender during the classifications. We also improve the state-of-the-art by presenting a comparison of other implementations of classification models. Besides cough, fever, sore throat, shortness of breath, and headache, we used the symptoms of olfactory disorders, coryza, and taste disorders to improve the results.

In contrast to such prior works, we focused on raw data from 55,676 Brazilians and used features that do not require expensive exams, such as CT images and blood tests. Symptoms included fever, sore throat, dyspnea, olfactory disorders, cough, coryza, taste disorders, and headache. The *Gender* and *Health Professional* features were the additional information required to conduct the COVID-19 test prioritization using the classification models. *Gender* was also used as a feature by prior works [4, 20, 28, 37]. The usage of exams such as CT images and blood tests limits classification models' application scenarios because it is necessary to prioritize symptomatic patients for testing in the first days after symptoms onset.

Limitations

By pre-processing the 55,676 raw data, the *RT-PCR Balanced* dataset only included 1,832 symptomatic patients, the *Rapid Balanced* dataset included 1,296 symptomatic patients, and the *Both Balanced* dataset included 3,128 symptomatic patients. However, to improve the strength of results and decrease size limitation, we also considered three unbalanced datasets. For example, the *Both Unbalanced* dataset is composed of 20,021 symptomatic patients tested for COVID-19.

Besides, in a real-world scenario, the number of asymptomatic patients with COVID-19 can also be considered a limitation to classification models' applicability. In this case, this study continues to be relevant due to the remaining symptomatic cases that also require healthcare professionals and the government's attention.. The evaluation of symptomatic patients is also relevant to prevent the unplanned usage of COVID-19 testing resources due to other disease outbreaks in Brazil caused by other viral infections (e.g., dengue, zika, and chikungunya). Such viral infections present similar symptoms that may difficult healthcare professionals' decision-making on the adequate choice about the needed testing types..

The reduced number of symptoms reported by a symptomatic patient can also negatively impact the reuse classification models. Nevertheless, the feature ranking and other information (e.g., contact with infected people) are relevant to complement the classification models during the decision-making conducted by healthcare professionals and policymakers. We verified the impacts of reducing features in the performance of implemented classification models (Multimedia Appendix

1).

Finally, the number of classification models implemented, validated, and compared is another limitation of our study, given the wide variety of available algorithms and ensemble strategies. This limitation was reduced by selecting well-known algorithms based on trees, linear regression, statistical learning, distance, and the concept of neurons.

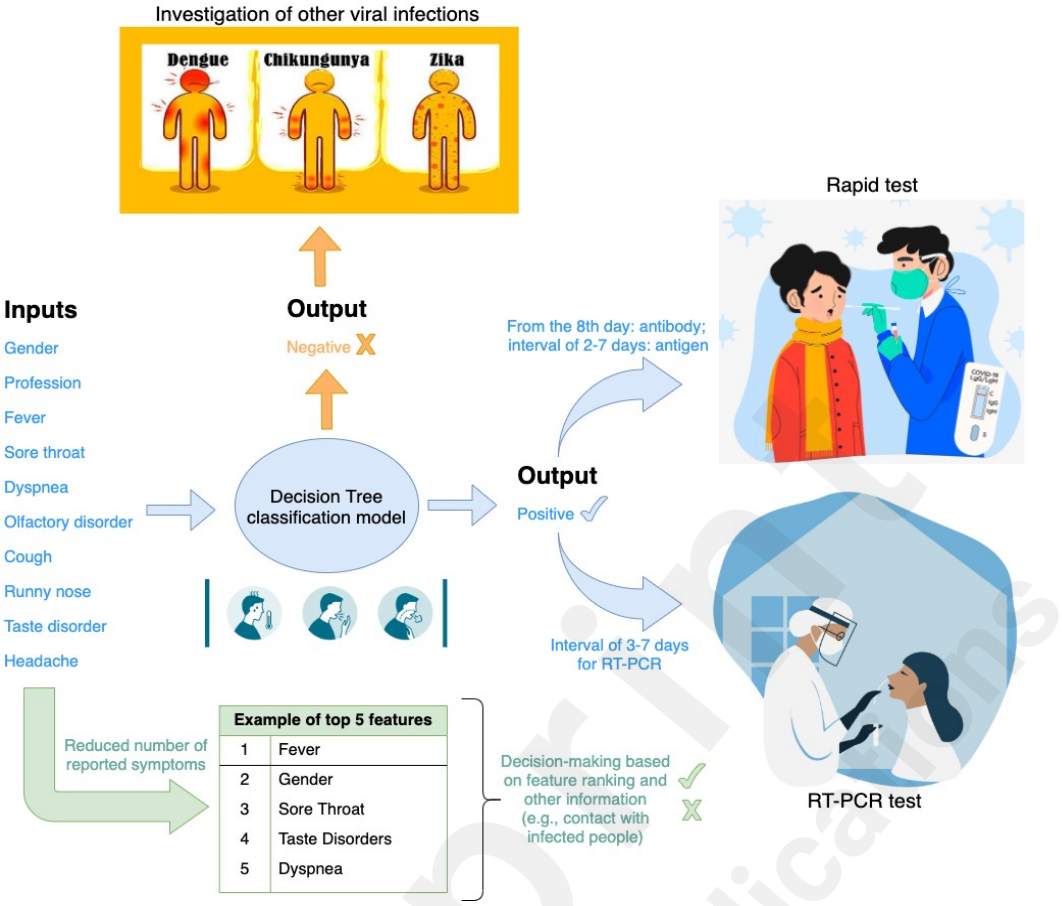
Clinical Practice Context

The availability of e-Health and m-Health systems is relevant to assist the decision-making in different scenarios. One such scenario is detecting COVID-19 in patients who reside in remote and hard-to-reach locations (e.g., Amazonia - Latin America) [38]. Developers can integrate e-Health and m-Health systems with services that enable healthcare professionals to be alerted when the risk of disease is detected. The usage of e-Health and m-Health systems should be encouraged, considering that the early detection of COVID-19 is essential in clinical practice to enable early medical attention, possibly reducing the negative impacts of late treatments. This type of e-Health and m-Health system can also benefit public health systems when factors related to the human condition (e.g., fatigue and lack of experience) and the collapse of health services negatively influence healthcare professionals' decision-making during patients' evaluation. Such scenarios are authentic in the context of the new coronavirus pandemic [39].

Therefore, the implemented classification models can be the basis for e-Health and m-Health systems to support healthcare professionals and policymakers during the COVID-19 test prioritization. To be applied in clinical practice and integrated with the current clinical workflow, it is recommended the availability of the DT classification model and the usage of feature ranking information through web services to be consumed by an e-Health or m-health system. Such a system shall present classification results for healthcare professionals in a user-friendly manner. The straightforward interpretation of classification models is relevant to increase healthcare professionals' confidence in classification results. For example, the web services can be integrated with Brazilian public health facilities systems to prioritize the reduced COVID-19 testing resources.

We present an application scenario integrating a clinical workflow and the DT classification model (Figure 7). The DT is used to prioritize symptomatic patients for COVID-19 testing. However, when the number of reported symptoms is too low, the classification models cannot distinguish between positive and negative cases. In this case, healthcare professionals can reuse the feature ranking and other information (e.g., contact with infected people) to make decisions about the COVID-19 testing. Thus, the usage of feature ranking information is guided by the answer of the SRQ 4. If the result is not prioritized, the patient's clinical condition should be further investigated in regards to other viral diseases.

Figure 7. An application scenario to connect the DT classification model with a clinical workflow. The model guides the test prioritization of patients suspected of COVID-19.



For the application scenario, there are five possible flows: (1) confirmed case with classification model, and rapid test result; (2) confirmed case with classification model, and RT-PCR test result; (3) confirmed case using feature ranking, and rapid test result; (4) confirmed case using feature ranking, and RT-PCR test result; and (5) negative case with the recommendation of investigation of other viral diseases. It is relevant to consider the days between the onset of symptoms and COVID-19 testing: closed interval of 3-7 days for RT-PCR test; from the 8th day for the rapid antibody test; and closed interval of 2-7 days for the rapid antigen test [40-42].

Conclusions

The results showed the relevance of using classification models for the COVID-19 test prioritization in Brazil, mainly based on the symptoms that do not require expensive exams. By comparing the classification models using raw data from 55,676 Brazilians, the 10-fold cross-validation method, classification metrics, and the Friedman and Nemenyi tests, the MLP, GBM, DT, RF, XGBoost, and SVM presented the highest performances with similar results.

Decision-tree-based classification models' high performances are relevant for our application scenario due to the high levels of decision trees' interpretability, positively impacting healthcare professionals' final decision-making. Therefore, applying the easy interpretability as an additional comparison criterion, DT was considered the most suitable classification model, effectively assisting in the decision-making for prioritizing symptomatic patients for testing. Information about the features *Gender*, *Health Professional*, *Fever*, *Sore Throat*, *Dyspnea*, *Olfactory Disorders*, *Cough*, *Coryza*, *Taste Disorders*, and *Headache* enable the COVID-19 test prioritization for symptomatic patients. The usage of symptoms that do not require expensive exams contributes to assisting patients who live, for example, in needy and hard-to-reach communities. The results of feature ranking reported in this article are also relevant to support a more detailed analysis in a scenario where a

patient reports a reduced number of symptoms.

To improve testing prioritization, we plan to investigate the relationship between the symptoms reported by COVID-19 patients and other widespread diseases in Brazil, such as dengue fever, Zika fever, and chikungunya. Thus, we aim to include implementing and validating classification models and developing and validating an e-health system to support healthcare professionals and policymakers in decision-making for testing strategies.

Acknowledgments

The authors thank the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil (CNPq), Federal University of the Agreste of Pernambuco, Coordenacão de Aperfeiçamento de Pessoal de Nível Superior - Brazil (CAPES), and Programa de Pós-Graduação em Engenharia Elétrica (COPELE), Federal University of Campina Grande.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Project 88881.507204/2020-01.

Conflicts of Interest

None declared.

Abbreviations

AP: Average Precision
AUROC: Area Under the Receiver Operating Characteristic Curve
CD: Critical Difference
CT: Computerized Tomography
DT: Decision Tree
ERT: Extremely Randomized Trees
GBM: Gradient Boosting Machine
KNN: K-Nearest Neighbor
LR: Logistic Regression
ML: Machine Learning
MLP: Multilayer Perceptron
NB: Naive Bayes
AUPRC: Area Under the Precision-Recall Curve
RF: Random Forest
RQ: Research Question
RT-PCR: Reverse Transcriptase-Polymerase Chain Reaction
SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
SD: Standard Deviation
SRQ: Secondary Research Question
SVM: Support Vector Machine
TWRF: Three-Way Random Forest
XGBoost: Extreme Gradient Boosting

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the dataset of Brazilian symptomatic patients for screening the risk of COVID-19, [10]. Besides, the source codes are available in a GitHub COVID-19 repository [12].

Authors' contributions

ÍVSS contributed to implementing the classification models, conducting the practical experimentation/validation of the models, and writing the article. AS, LCS, LDS, ECG, and AP contributed to the study's conception, research methodology, revision of implementation and validation results, and the article's writing and revisions. ACMS and DFSS contributed to the revisions of implementation, validation results, and article.

Multimedia Appendix 1

Performance of classification models considering the Chi-squared results.

Multimedia Appendix 2

Results of statistical tests.

Multimedia Appendix 3

Feature ranking results for the unbalanced dataset.

Multimedia Appendix 4

Mean values of classification metrics for the ensemble models.

References

1. Belard A, Buchman T, Forsberg J, et al. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. J Clin Monit Comput. 2017;31(2):261-271. doi:10.1007/s10877-016-9849-1
2. Elhoseny M, Abdelaziz A, Salama AS, et al. A hybrid model of internet of things and cloud computing to manage big data in health services applications. Future Gener Comput Syst. 2018; 86:1383-1394. doi:10.1016/j.future.2018.03.005
3. Chatterjee A, Gerdes MW, Martinez S. eHealth Initiatives for The Promotion of Healthy Lifestyle and Allied Implementation Difficulties. Conference Proceedings of the International Conference on Wireless and Mobile Computing, Networking and Communications. October 2019; 1-8. Barcelona, Spain. doi: 10.1109/WiMOB.2019.8923324
4. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digit Med. 2021;4(1):3. Published 2021 Jan 4. doi:10.1038/s41746-020-00372-6
5. Guimarães VHA, de Oliveira-Leandro M, Cassiano C, et al. Knowledge About COVID-19 in Brazil: Cross-Sectional Web-Based Study. JMIR Public Health Surveill. 2021;7(1):e24756. Published 2021 Jan 21. doi:10.2196/24756
6. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5
7. Veiga E Silva L, de Andrade Abi Harb MDP, Teixeira Barbosa Dos Santos AM, et al. COVID-19 Mortality Underreporting in Brazil: Analysis of Data From Government Internet Portals. J Med Internet Res. 2020;22(8):e21413. Published 2020 Aug 18. doi:10.2196/21413
8. Ferrante L, Steinmetz WA, Almeida ACL, et al. Brazil's policies condemn Amazonia to a second wave of COVID-19. Nat Med. 2020;26(9):1315. doi:10.1038/s41591-020-1026-x
9. Monteiro de Oliveira M, Fuller TL, Brasil P, Gabaglia CR, Nielsen-Saines K. Controlling the COVID-19 pandemic in Brazil: a challenge of continental proportions. Nat Med. 2020;26(10):1505-1506. doi:10.1038/s41591-020-1071-5
10. Santana IVS, Sobrinho A, Silva LC, et al. Brazilian dataset of symptomatic patients for

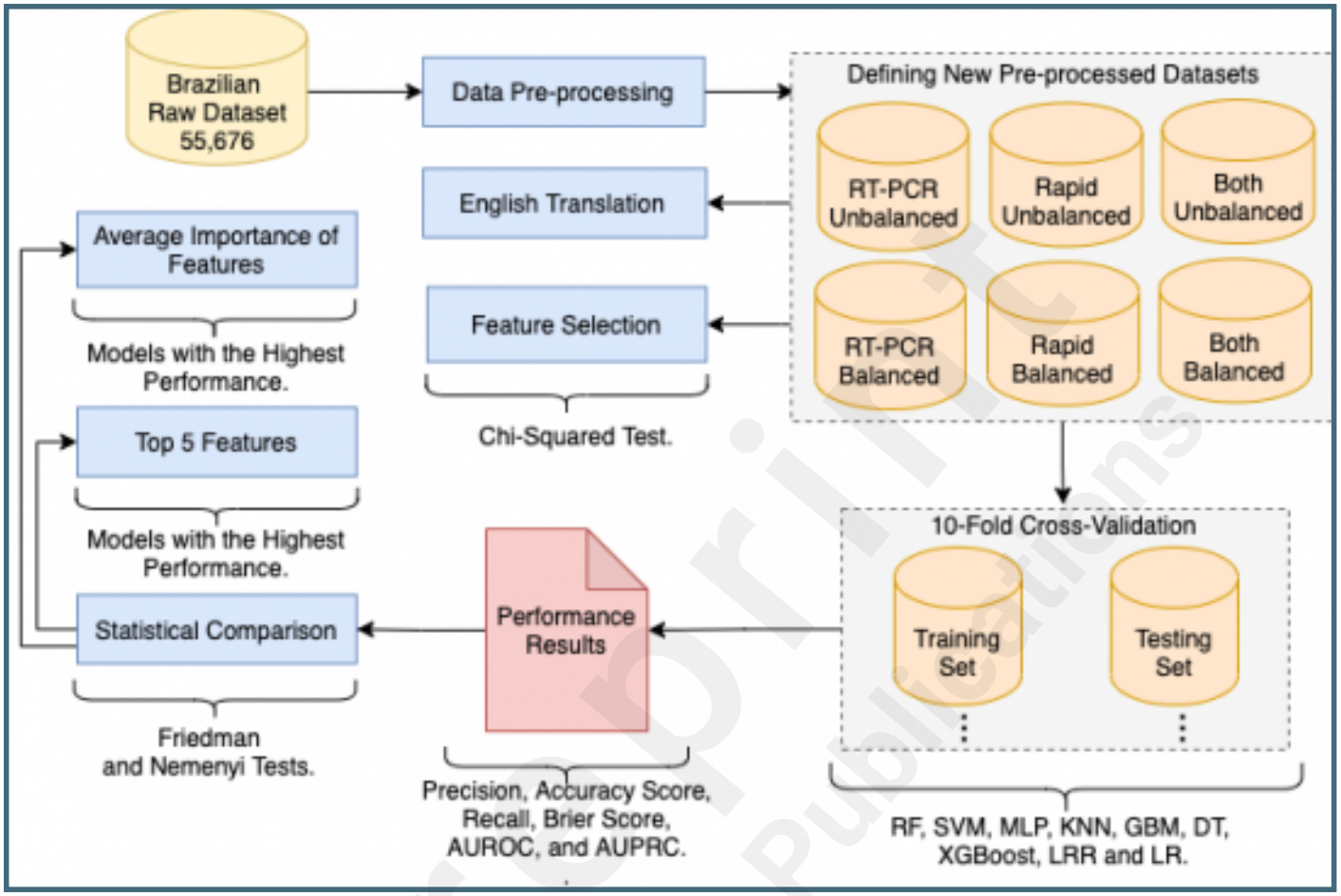
screening the risk of COVID-19. 2021. doi:10.17632/b7zcgmmwx4.5
11. Malta M, Murray L, da Silva CMFP, Strathdee SA. Coronavirus in Brazil: The heavy weight of inequality and unsound leadership. EClinicalMedicine. 2020;25:100472. doi:10.1016/j.eclinm.2020.100472
12. Santana IVS, Sobrinho A, Silva LC, et al. GitHub COVID-19 repository 2021. URL: http://bit.ly/2Met3S4
13. Chatterjee A, Gerdes MW, Martinez SG. Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview. Sensors (Basel). 2020;20(9):2734. doi:10.3390/s20092734
14. World Health Organization. Gender and COVID-19. Advocacy brief 2020.
15. Peckham H, de Gruijter NM, Raine C, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun. 2020;11(1):6317. doi:10.1038/s41467-020-19741-6
16. Shilaskar S, Ghatol A. Diagnosis system for imbalanced multi-minority medical dataset. Soft Comput 2019; 23:4789–4799. doi:10.1007/s00500-018-3133-x
17. Almansour NA, Syed HF, Khayat NR, et al. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Comput Biol Med. 2019;109:101-111. doi:10.1016/j.compbiomed.2019.04.017
18. Biau G, Cadre B, Rouviere L. Accelerated gradient boosting. Mach Learn. 2019;108:971-992. doi:10.1007/s10994-019-05787-1
19. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451
20. Xing W, Bei Y. Medical health big data classification based on knn classification algorithm. IEEE Access. 2020;8:28808-28819. doi:10.1109/ACCESS.2019.2955754
21. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl. 2020;160:113661. doi:10.1016/j.eswa.2020.113661
22. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. 2020;121:103761. doi:10.1016/j.compbiomed.2020.103761
23. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci. 2017;9:329. doi:10.3389/fnagi.2017.00329
24. Schober P, Vetter TR. Logistic Regression in Medical Research. Anesth Analg. 2021;132(2):365-366. doi:10.1213/ANE.0000000000005247
25. Scikit-learn: Machine Learning in Python. URL: https://scikit-learn.org/stable/. [accessed 2021-02-21]
26. Ferri C, Hernandez-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognit Lett. 2009;30(1):27-38. doi:10.1016/j.patrec.2008.08.010
27. Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.
28. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020;26(8):1224-1228. doi:10.1038/s41591-020-0931-3
29. The Lancet Respiratory Medicine. Opening the black box of machine learning. Lancet Respir Med. 2018;6(11):801. doi:10.1016/S2213-2600(18)30425-9
30. Centers for Disease Control and Prevention: Is it Dengue or is it COVID-19? 2020. URL: http://bit.ly/35u3t2p [accessed 2021-02-21]
31. Son WS, Park JE, Kwon O. Early detection of influenza outbreak using time derivative of incidence. EPJ Data Sci. 2020;9(1):28. doi:10.1140/epjds/s13688-020-00246-7

32. Chatterjee A, Gerdes MW, Martinez SG. Statistical Explorations and Univariate Timeseries Analysis on COVID-19 Datasets to Understand the Trend of Disease Spreading and Death. Sensors (Basel). 2020;20(11):3089. doi:10.3390/s20113089
33. Kasson PM. Infectious disease research in the era of big data. Annu Rev Biomed Data Sci. 2020;3(1):43-59. doi:10.1146/annurev-biodatasci-121219-025722
34. Kumar S. Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. Ann Data Sci. 2020;7:417-425. doi:10.1007/s40745-020-00289-7
35. Khakharia A, Shah V, Jain S, et al. Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning. Ann Data Sci. 2020;22(11):24018. doi:10.2196/24018
36. Vaid A, Somani S, Russak AJ, et al. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. J Med Internet Res. 2020;22(11):e24018. doi:10.2196/24018
37. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst. 2020;44(8):135. Published 2020 Jul 1. doi:10.1007/s10916-020-01597-4
38. Sobrinho A, da Silva LD, Perkusich A, Pinheiro ME, Cunha P. Design and evaluation of a mobile application to assist the self-monitoring of the chronic kidney disease in developing countries. BMC Med Inform Decis Mak. 2018;18(1):7. doi:10.1186/s12911-018-0587-9
39. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med. 2020;26(4):459-461. doi:10.1038/s41591-020-0824-5
40. Lima FET, Albuquerque NLS, Florencio SSG, et al. Time interval between onset of symptoms and COVID-19 testing in Brazilian state capitals. Epidemiol Serv Saude. 2020;30(1):e2020788. doi:10.1590/S1679-4974202100010002
41. Zhao J, Yuan Q, Wang H, et al. Antibody Responses to SARS-CoV-2 in Patients With Novel Coronavirus Disease 2019. Clin Infect Dis. 2020;71(16):2027-2034. doi:10.1093/cid/ciaa344
42. Burki TK. Testing for COVID-19. Lancet Respir Med. 2020;8(7):e63-e64. doi:10.1016/S2213-2600(20)30247-2

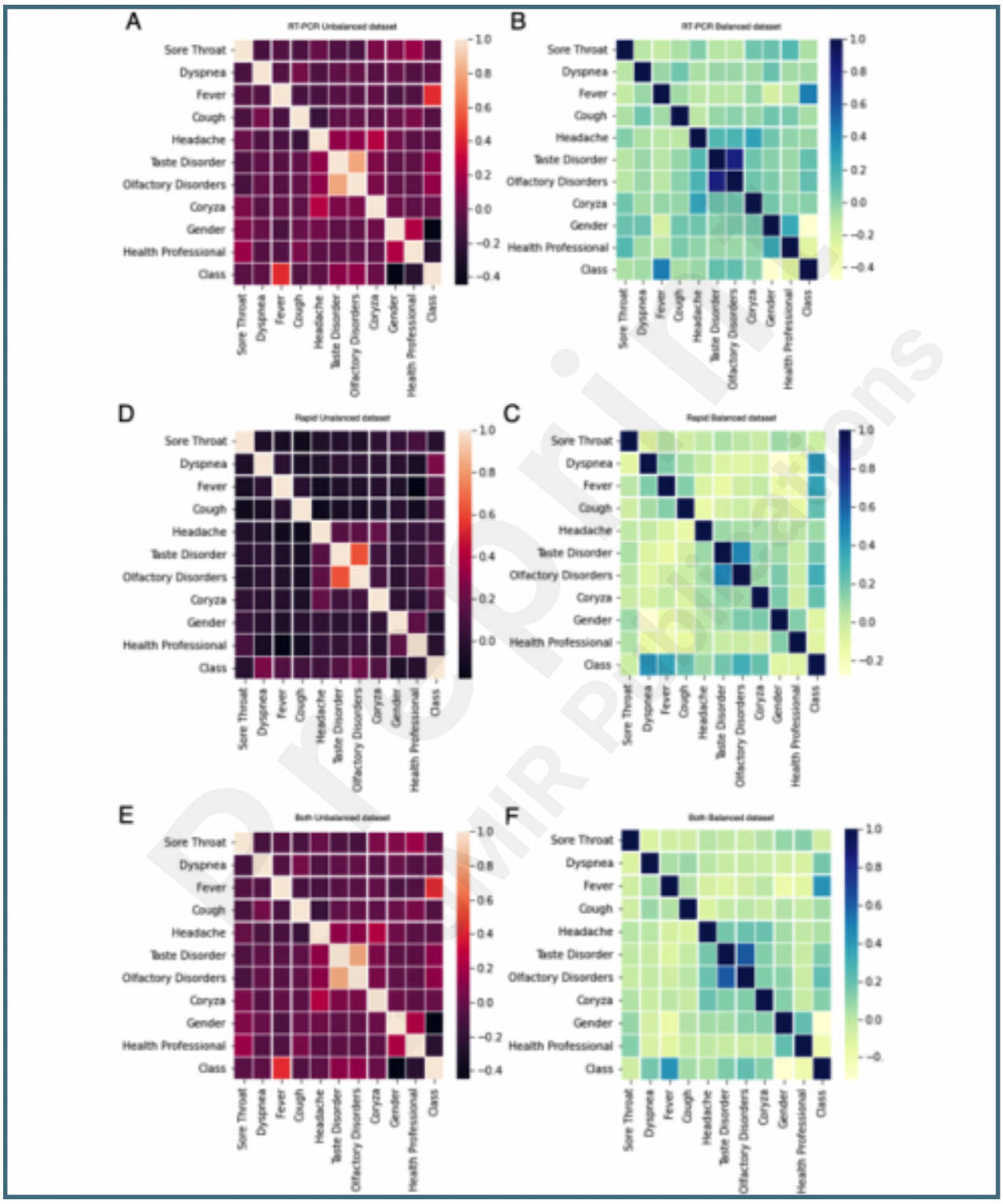
Supplementary Files

Figures

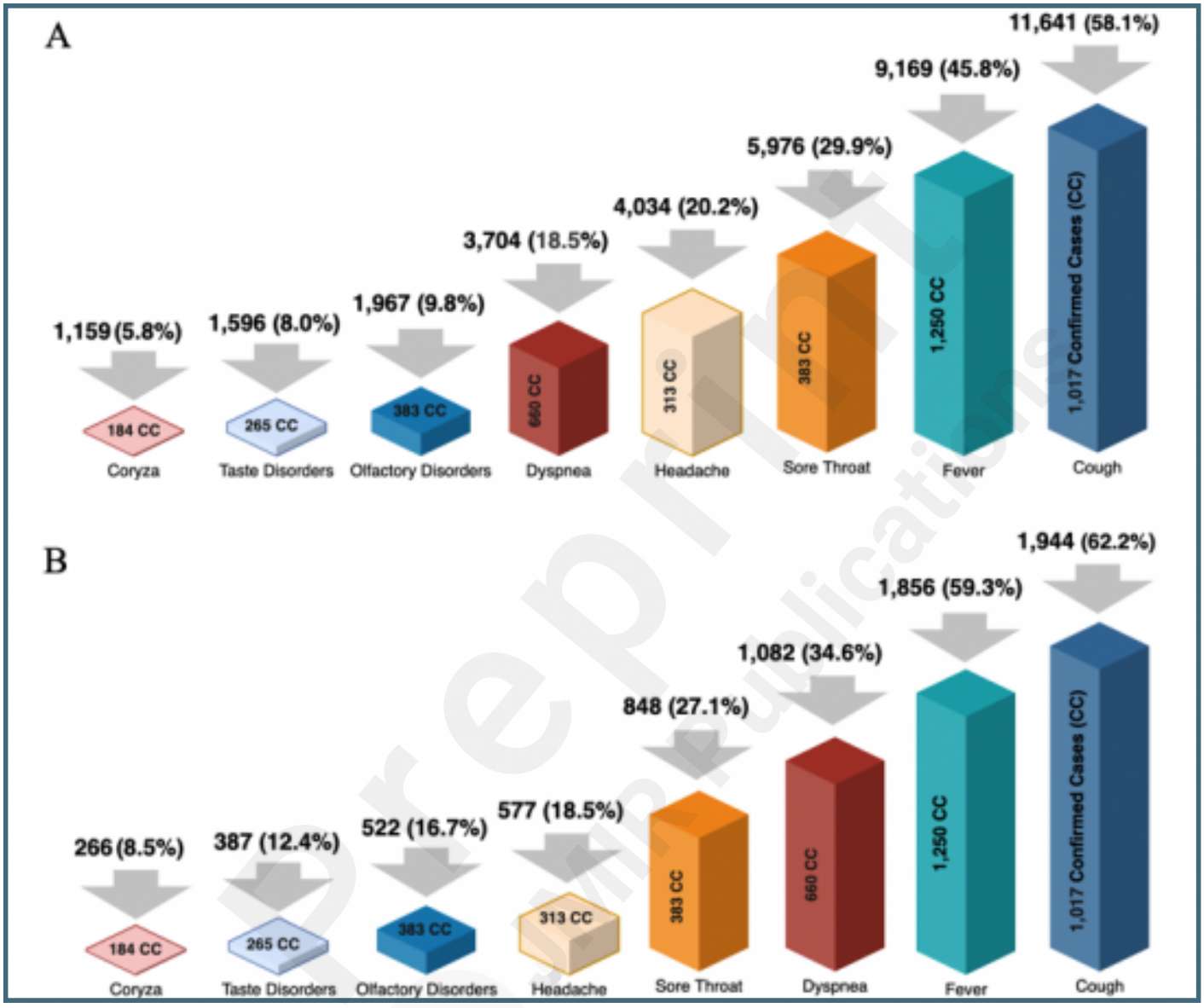
Overview of the research methodology applied for the study. The methodological steps consist of data pre-processing, the definition of new datasets, English translation, feature selection, 10-fold cross-validation, statistical comparisons, and feature ranking.



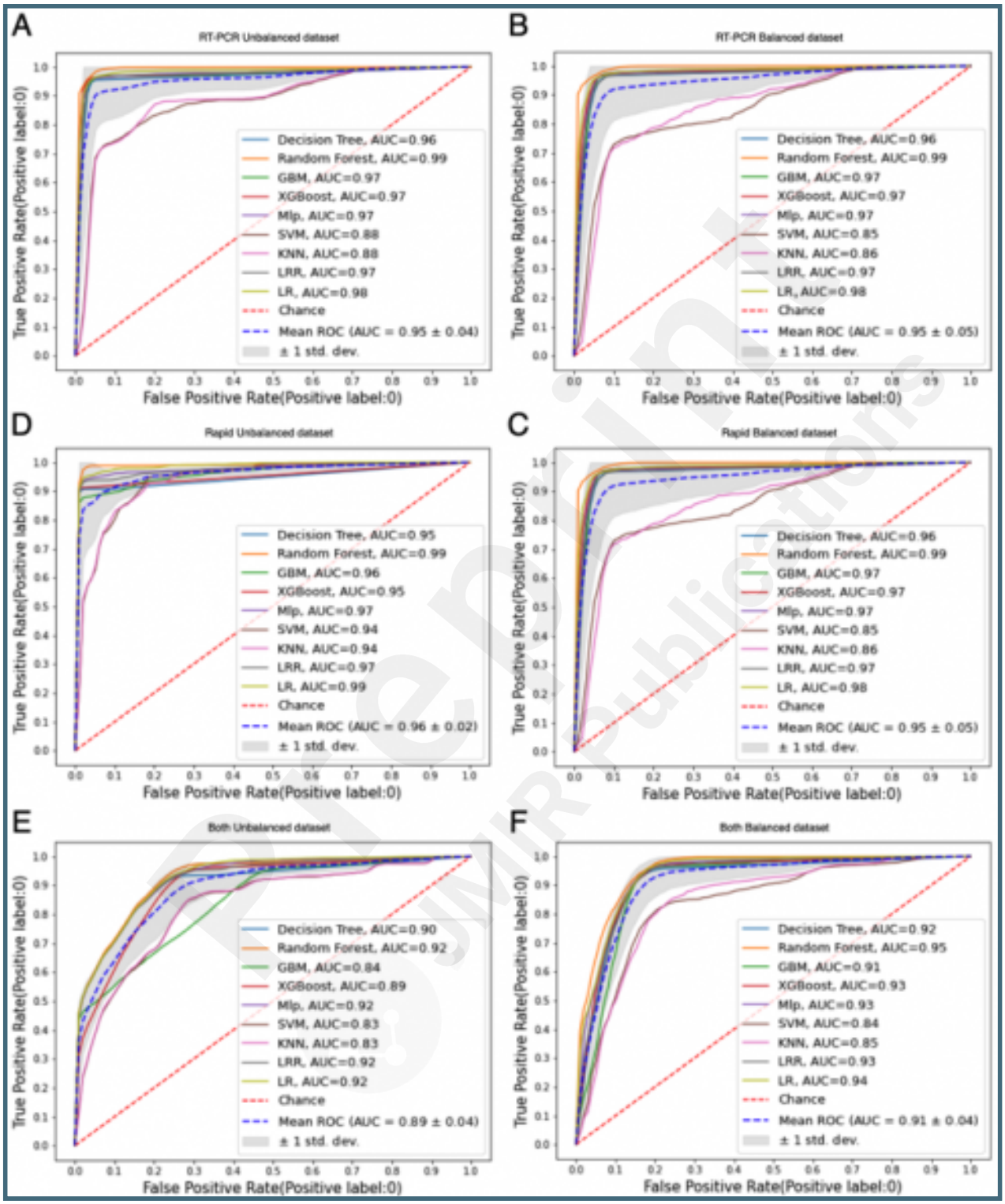
Correlation matrix for (A) RT-PCR Unbalanced dataset, (B) RT-PCR Balanced dataset, (C) Rapid Unbalanced dataset, (D) Rapid Balanced dataset, (E) Both Unbalanced dataset, and (F) Both Balanced dataset.



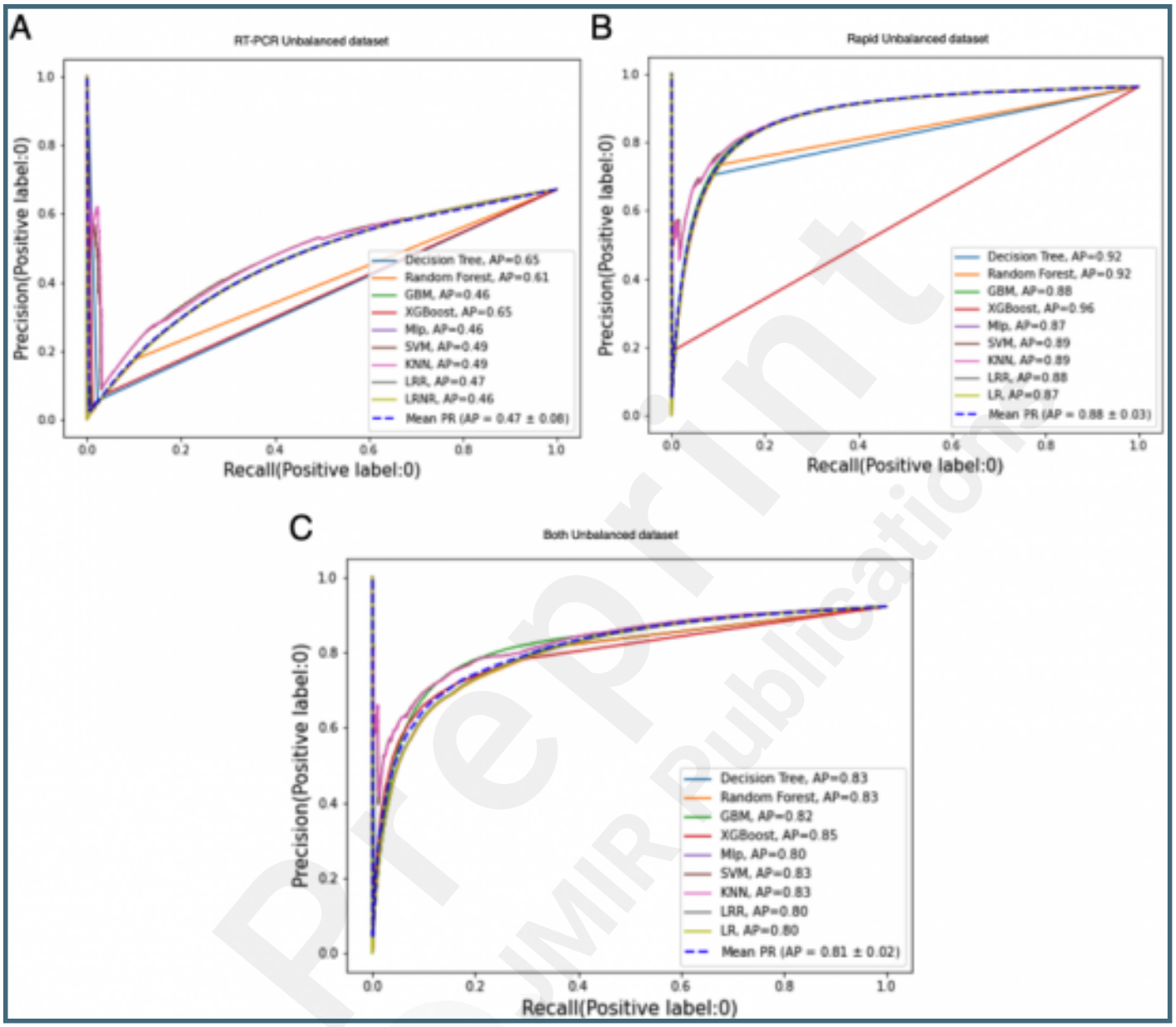
(A) The frequency of symptoms for the 20,021 symptomatic patients of the Both Unbalanced dataset and the number of Confirmed Cases (CC). Top values are frequencies; numbers on the geometric forms are the CC for frequency. (B) The frequency of symptoms for the 3,128 symptomatic patients of the Both Balanced dataset and the number of CC.



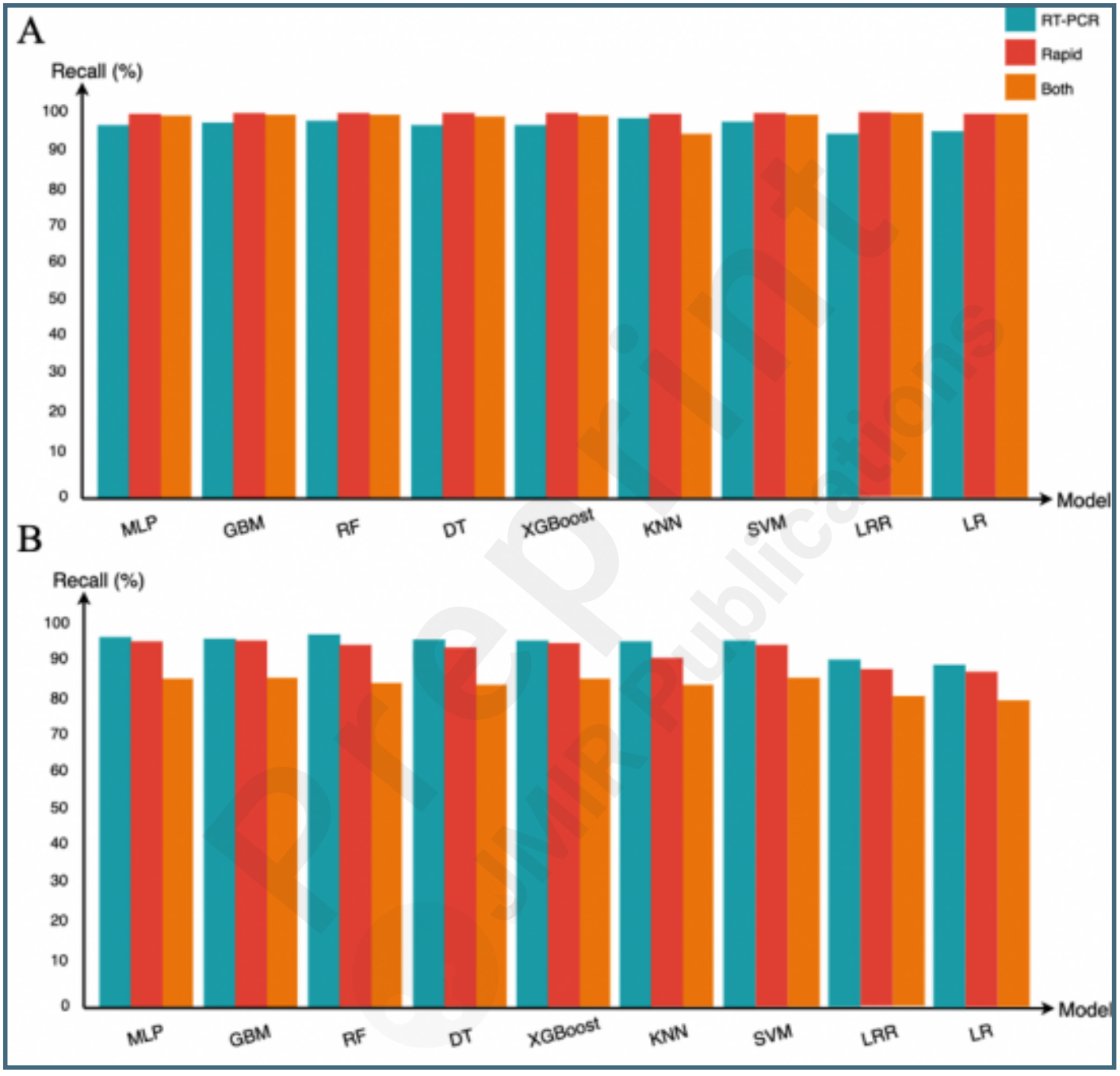
Models' ROC curves with (A) RT-PCR Unbalanced, (B) RT-PCR Balanced, (C) Rapid Unbalanced, (D) Rapid Balanced, (E) Both Unbalanced, and (F) Both Balanced.



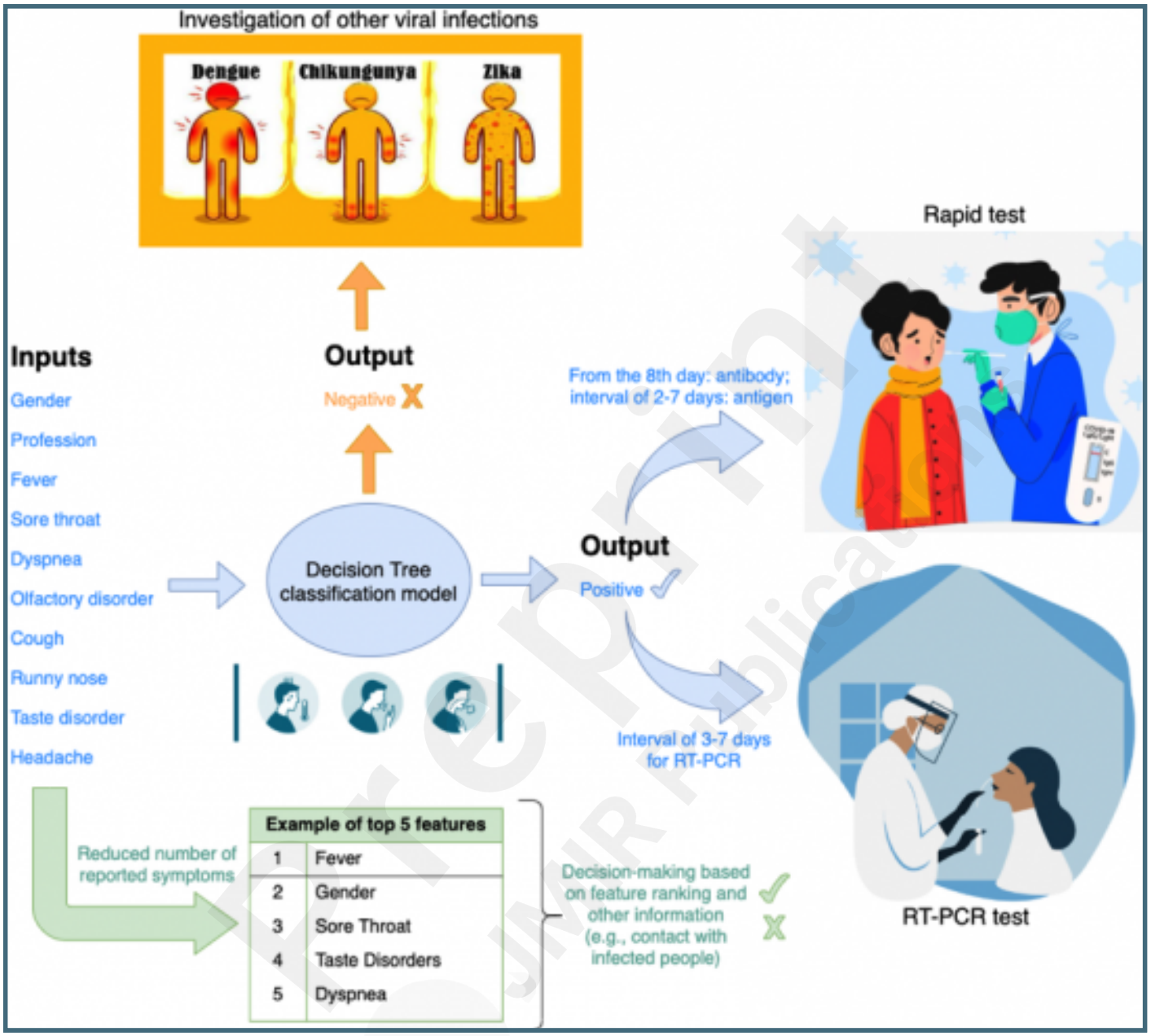
Models' PRC with (A) RT-PCR Unbalanced dataset, (B) Rapid Unbalanced dataset, and (C) Both Unbalanced dataset.



(A) The mean recall for the MLP, GBM, RF, DT, XGBoost, KNN, SVM, LRR, and LR classification models using the unbalanced datasets for RT-PCR, rapid, and both types. (B) The mean recall for the MLP, GBM, RF, DT, XGBoost, KNN, SVM, LRR, and LR classification models using the balanced datasets for RT-PCR, rapid, and both types.



An application scenario to connect the DT classification model with a clinical workflow. The model guides the test prioritization of symptomatic patients suspected of COVID-19.



Multimedia Appendixes

Performance of classification models considering the Chi-squared results.
URL: http://asset.jmir.pub/assets/6a22fa3f73ce2798bca6ce5d92b1f29f.docx

Results of statistical tests.
URL: http://asset.jmir.pub/assets/8da37ba4503e74ba14076b7304cb0bdb.docx

Feature ranking results for the unbalanced dataset.
URL: http://asset.jmir.pub/assets/cb049ac9bf389570146ece93d909776c.docx

Mean values of classification metrics for the ensemble models.
URL: http://asset.jmir.pub/assets/9a9e00ac8283081c62f79793adfb2756.docx