High performance COVID-19 screening using machine learning

Dépistage de haute performance du COVID-19 utilisant l'apprentissage machine

Youssef Zied Elhechmi[1], Mehdi Mrad[2], Mariem Gdoura[3], Anissa Nouri[4], Helmi Ben Saad[5], Najla Ghrairi[1], Henda Triki[2]

1. Hope Horizon International
2. Laboratory of viruses, vectors and hosts: LR20IPT10, Institut Pasteur de Tunis, University of Tunis El Manar, 13, Place Pasteur, 1002 Tunis-Belvédère, Tunisia
3. Laboratory of Clinical Virology, Institut Pasteur de Tunis, University of Tunis El Manar, 13, Place Pasteur, 1002 Tunis-Belvédère, Tunisia.
4. Clinical investigation center: 2016CICIPT02, Institut Pasteur de Tunis, University of Tunis El Manar, 13, Place Pasteur, 1002 Tunis Belvédère, Tunisia.
5. MD (Faculty of Medicine of Sousse, Tunisia). PhD (Faculty of Medicine of Montpellier, France). Physiology and Functional Explorations. Laboratory of Physiology and Functionnal Explorations. Farhat Havhed Hospital. Sousse, Tunisia. Laboratory of Physiology. Faculty of Medicine of Sousse. Street Mohamed Karoui. Sousse 4000. Tunisia.

Abstract

Since the World Health Organization declared the Coronavirus Disease 2019 (COVID-19) pandemic as an international concern of public health emergency in the early 2020, several strategies have been initiated in many countries to prevent healthcare services breakdown and collapse of healthcare structures. The most important strategy was the increased testing, diagnosis, isolation, contact tracing to identify, quarantine and test close contacts. In this context, finding a rapid, reliable and affordable tool for COVID-19 screening was the main challenge to address the pandemic. Molecular diagnosis by reverse transcriptase polymerase chain reaction (RT-PCR), even though considered as the gold standard in the diagnosis of COVID-19, was time consuming and therefore does not fit the objective of rapid screening. In addition, serological tests to detect anti-severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) antibodies suffered from low sensitivity. Prediction models based on machine-learning (ML) that combined several clinical features to estimate the risk of COVID-19 have been developed. To address these screening challenges, we created a ML model (MLM) based on gradient boosting method. We included several clinical features and the daily geographic prevalence of COVID-19 cases in the MLM. The MLM was trained on 1554 cases (757 COVID-19), and tested on 547 cases (169 COVID-19). Our MLM successfully predicted RT-PCR positivity with an accuracy of 97.06%. Moreover, the variable sensitivity and specificity of our MLM depending on the disease geographic prevalence has introduced the concept of “dynamic” disease screening. In the context of future world pandemic emergencies, we believe that this MLM method can be very useful as a rapid, reliable and dynamic screening tool for contagious diseases, especially in the developing countries.

Key words: Artificial intelligence, COVID-19, Machine learning, Mass screening, Public health.

Résumé

Depuis que l’Organisation mondiale de la santé a déclaré la pandémie de coronavirus (COVID-19) urgence de santé publique en 2020, plusieurs stratégies ont été lancées pour prévenir l’effondrement des structures de santé. La stratégie la plus importante a été l’augmentation du dépistage, diagnostic, isolement, et recherche des contacts pour identifier, mettre en quarantaine et tester les contacts étroits. Dans ce contexte, la principale difficulté était de trouver un outil rapide, fiable et abordable pour le dépistage de la COVID-19. Le diagnostic moléculaire par "polymerase chaine reaction" (RT-PCR), bien que considéré comme le "gold-standard", était lent et ne correspond donc pas à l’objectif du dépistage rapide. En outre, les tests sérologiques visant à détecter les anticorps anti-coronavirus 2 souffraient d’une faible sensibilité. Pour relever ces défis, nous avons créé un modèle d'apprentissage machine basé sur la méthode du "gradient boosting". Nous avons inclus plusieurs caractéristiques cliniques avec la prévalence géographique quotidienne des cas COVID-19 dans le modèle. Le modèle a été entrainé sur 1554 cas (757 COVID-19) et testé sur 547 cas (169 COVID-19). Notre modèle a prédit la positivité de RT-PCR avec une précision de 97,06%. De plus, la sensibilité et la spécificité variables du modèle dépendant de la prévalence géographique de la maladie ont introduit le concept du dépistage « dynamique » . Dans le contexte des futures urgences pandémiques mondiales, nous pensons que cette méthode peut être très utile comme outil de dépistage rapide, fiable et dynamique pour les maladies contagieuses, en particulier dans les pays en développement.

Mots clés: Intelligence artificielle, COVID-19, Apprentissage, Apprentissage machine, Santé publique

Correspondance
Youssef Zied Elhechmi
Hope Horizon International
Email: youssef@hopehorizonworld.com

LA TUNISIE MEDICALE-2025; Vol 103 (01): 10-17 DOI: 10.62438/tunismed.v103i1.5401

Introduction

In December 2019, a previously unknown Beta-coronavirus was discovered in a cluster of patients hospitalized for pneumonia in Wuhan, China [8]. Since then, the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) spread to the rest of the world in what it was called the Coronavirus Disease 2019 (COVID-19) pandemic. To fight against this pandemic, several strategies have been initiated (eg; physical distancing, limiting contacts, avoiding nonessential indoor spaces and crowded outdoor settings, safeguarding persons most at risk for severe illness or death, protecting essential exposed workers, postponing travel, increased room air ventilation, enhanced hand hygiene, cleaning and disinfecting, widespread availability and use of effective vaccines. However, in terms of rapid prevention, the most important strategy may be the increased testing, diagnosis, isolation, contact tracing to identify, quarantine and test close contacts [4]. To reach this objective, the gold standard is the reverse transcription polymerase chain reaction (RT-PCR), however this technique takes several hours to confirm positivity and this time span is too long [9]. While waiting up to 7 days for RT-PCR confirmation of SARS-COV-2 infection, patients may continue to infect other people [9]. Because of contagiousness considerations, patients who consult emergency departments cannot be dispatched to hospital wards before getting a RT-PCR result confirming or denying their infection and this raises the issue of emergency department overcrowding and the risk of hospital contamination of patients. Wang et al. [10] reported that among 138 hospitalized patients with COVID-19, 41% were suspected to be infected via hospital related transmission. On the other hand, rapid testing for SARS-COV-2 methods showed a low sensitivity in some studies and a negative test might provide false reassurance [11]. In this context, the growing interest of the applications of artificial intelligence in medicine is strongly justified [12]. Several studies have been conducted to develop COVID-19 diagnosis systems using artificial intelligence (AI) techniques [13–16]. Including chest tomography (CT) scan, biological or clinical data, all these studies focused on the patient's symptoms and explorations and proved a high performance in the prediction of COVID-19 diagnosis [16–21]. We aimed to create a machine-learning model (MLM) using clinical features along with geographic COVID-19 prevalence for the prediction of SARS-Cov-2 RT-PCR positivity.

Methods

Study design

This was a prospective observational study including all patients screened for COVID-19 using RT-PCR performed in the Virology laboratory of the Pasteur Institute of Tunis from March 2020 to October 2020, covering the first and the ascension of the second wave of COVID-19 pandemic in Tunisia. This study obtained the consent of the Ethics Committee of the Pasteur Institute of Tunis.

Population

Patients with clinical suspicion of COVID-19 were selected in all the 24 governorates of Tunisia.

Applied protocol

COVID-19 clinical suspicion was based on the presence of at least one symptom evoking COVID-19. Patients were sampled using a nasopharyngeal swab, and a sheet was carefully filled by a physician including fever, asthenia, anosmia, ageusia, headache, myalgia, throat pain, dyspnea, cough, rhinitis, chest pain, diarrhea, vomiting, nausea, abdominal pain. The sheet was, then, sent to the virology lab of the Pasteur Institute with the samples according to the World Health Organization (WHO) guidelines for packaging and shipment related to SARS-COV-2 [22]. Viral RNA was extracted from 200 µl of nasopharyngeal swabs using the QIAamp viral RNA Mini Kit (Qiagen, Hilden, Germany). Viral RNA was subsequently amplified and detected using a real-time fluorescent RT-PCR in house assay approved by the WHO: the Hong Kong university protocol targeting two genome regions: N and Orf 1b nsp 14 [23]. The N gene RT-PCR is recommended as a screening assay and the Orf1b assay as a confirmatory one [23]. Both should give cycle threshold (Ct) value < 40 to consider the sample as positive [23]. Suitable biosafety precautions were taken for handling human clinical specimens suspected to be SARS-COV-2 infections [24].

Daily incidence (active cases/100'000 inhabitants) of new COVID-19 cases by governorate were collected from the public data communicated by the Tunisian Health Ministry.

All the parameters mentioned in the study form were included in the statistical analysis and in the model training in addition to the COVID-19 incidence. A statistical analysis was conducted to look for parameters correlated with the RT-PCR positivity using Epi info software (CDC, USA) [25]. We selected 75% of the cases to train a MLM based on extreme gradient boosting method, and 25% of the cases were used to test the MLM and assess its accuracy for the prediction of positive RT-PCR. "p" was considered positive when ≤ 0.05.

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. Gradient boosting is a method of converting weak learners into strong learners. It trains several models in a gradual, additive and sequential manner and identify the shortcomings by using gradient descent optimization algorithm in the loss function. The latter is a measure indicating how good the model's coefficient are at fitting the data to predict. A definitive objective of the gradient boosting method is to discover a function (ie; F(x)), which limits its loss function "L" (ie; y, F(x)), through iterative back-fitting as

$$F^* = \underset{F}{\operatorname{argmin}} E_{y,x} L(y,F(x))$$

F*: This represents the optimal model that we're trying to find, which minimizes the error (loss) between the predicted values and the actual values.
arg min$_F$: Represents the function that gives the smallest predicted value of the expression following it. This means we have to find a predicted value/F for which the loss function is minimum.
E$_{y,x}$: This denotes the expected value over the joint distribution of the true target values y and the input features x.
L(y,F(x)): This is the loss function (or error function). It measures how far off the predictions F(x) are from the actual values y. Common loss functions include Mean Squared Error (for regression) or Log Loss (for classification).
In gradient boosting, the goal is to find a model F(x) that minimizes the loss function L(y,F(x)), where y represents the true values and F(x) represents the predicted values. The process of minimizing this error involves iterative steps where at each stage, the algorithm adds a new "weak learner" to correct the errors made by the previous model.

By definition, a supported predicted model is a weighted straight of the base learners

$$\mathbf{F}\left(\mathbf{x};\{\mathbf{B_m},\mathbf{A_m}\}_{\mathbf{1}}^{\mathbf{M}}\right)=\sum_{\mathbf{m=1}}^{\mathbf{M}}\mathbf{B_m p(x:a_m)}$$

Where p (x; a) is a base learners parameter [14,26].
Extreme gradient boosting (XGBoost) is an efficient open-source implementation of the gradient boosting algorithm [1]. Accordingly, the two main reasons to use XGBoost are execution speed and model performance. The XGBoost hyperparameters tuned in this study included Eta, max depth, min child weight, subsample, gamma and reg alpha. The hyperparameters where tuned to make balance between high performance of the model and low risk of over-fitting.
Eta = 0.56: Makes the model more robust by shrinking the weights on each step
Max depth = 6: Represents the maximum depth of a tree. Used to control over-fitting, as higher depth will allow model to learn relations very specific to a particular sample. Typical values range from 3 to 10.
Reg Alpha = 1.2: Represents the L1 regularization term on weight (analogous to Lasso regression). Can be used in case of very high dimensionality so that the algorithm runs faster when implemented.
Min child weight = 0: Defines the minimum sum of weights of all observations required in a child. This refers to min "sum of weights" of observations. Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting. Default value is 1.
Subsample = 0.2: Denotes the fraction of observations to be randomly samples for each tree. Lower values make the algorithm more conservative and prevents overfitting but too small values might lead to under-fitting. Typical values range from 0.5 to 1.
Gamma = 0: A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split. Makes the algorithm conservative. The values can vary depending on the loss function and is recommended to be tuned.

Python code

```
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
X_train, X_test, y_train, y_test=train_test_split(X,y,test_
size=0.25, random_state=0)
classifier=XGBClassifier(eta=0.56, max_depth=6, reg_
alpha=1.2,
min_child_weight=0,subsample=0.2,gamma=0)
classifier.fit(X_train, y_train)
```

The loss function used to train the model was the Mean Squared Error (MSE) which is the default loss function of the extreme gradient boosting algorithm.

$$\boldsymbol{MSE}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=0}}^{\boldsymbol{n}}\left(\boldsymbol{y_i}-\boldsymbol{y_{pred}}\right)^{\mathbf{2}}$$

Cross validation process was performed using the area under the receiver operating curve as metrics:

```
params = {"objective":"binary:logistic",'eta':0.56, 'max_
depth':6, 'reg_alpha':1.2,
        'min_child_weight':0, 'subsample': 0.3,'gamma': 0}
xgb_cv = cv(dtrain=dataMatrix, params=params,
nfold=3,
        num_boost_round=50, early_stopping_rounds=10,
metrics="auc", as_pandas=True, seed=123)
```

To estimate the prediction accuracy of the MLM, we used the area under the receiver operating characteristics curve (AURCC). A receiver operating characteristics curve is a graph representing the performance of a classification model for all classification thresholds. The curve plots the rate of true positives based on the rate of false positives. Different model's probability thresholds (eg; 25%, 50%, 75%) were then evaluated using sensitivity (recall), specificity (selectivity), positive (precision) and negative predictive values, and the Matthews correlation coefficient (MCC).

$$\boldsymbol{Sensitivity}=\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

$$\boldsymbol{Specificity}=\frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}}$$

$$\boldsymbol{Positive\ Predictive\ Value}=\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

$$\boldsymbol{Negative\ Predictive\ Value}=\frac{\boldsymbol{TN}}{\boldsymbol{TN+FN}}$$

Where TP is true positive; TN is true negative; FP is false positive; and FN is false negative.
The MCC is used in machine learning as a measure of the quality of binary (two-class) classifications [2]. The MCC ranges from -1 to 1. It takes into account TP, TN, FP, and

FN, and is generally regarded as a balanced measure, which can be used even if the classes are of very different sizes [3]. The closer the value is to 1, the more powerful the model is. The closer the value is to 0, the lower the power of the model.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Results

Between March and October 2020, 1884 samples were screened by RT-PCR for SARS-COV-2, 928 (49.25%) were positive and 956 (50.74%) were negative. Data about the prevalence of active cases per 100'000 inhabitants during the period of the study were collected from the Tunisian national health ministry (Figure 1). There was no missing data in this study.

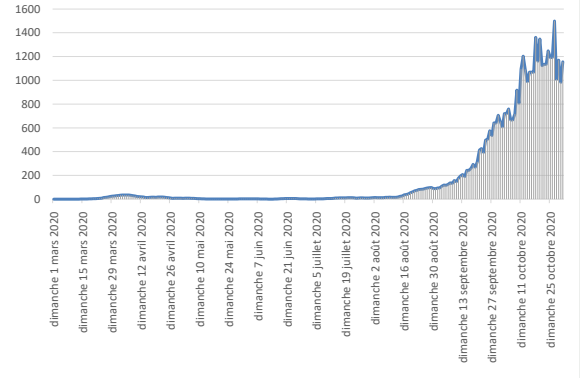

Figure 1. Prevalence of active cases of COVID-19 (per 100'000 inhabitants) from March 1st to October 31st, 2020.

A statistically significant association was found between positive RT-PCR test and fever, asthenia, anosmia, ageusia, headache, myalgia, cough, dyspnea, chest pain, diarrhea, vomiting (Table1).

Table 1. Correlation between clinical symptoms and reverse transcription polymerase chain reaction test for SARS-Cov-2.

Clinical feature	Negative RT-PCR[1] (n=956)	Positive RT-PCR[1] (n=928)	p value
Fever	24.50%	67.30%	<0.001
Asthenia	13.10%	32.20%	<0.001
Anosmia	0.40%	15.30%	<0.001
Ageusia	0.10%	8.90%	<0.001
Headache	13.20%	25.20%	<0.001
Myalgia	9.20%	29.20%	<0.001
Sore throat	4.30%	5.30%	0.313
Cough	35.30%	59.50%	<0.001
Dyspnea	16.10%	36.30%	<0.001
Rhinitis	10.90%	13.10%	0.130
Chest pain	4.70%	7.10%	0.027
Diarrhea	6.70%	14.00%	<0.001
Vomiting	3.80%	6.90%	0.002
Abdominal pain	4.10%	5.80%	0.081

[1] SARS-Cov-2 RT-PCR

The prevalence of COVID-19 active cases was significantly associated with the RT-PCR positivity. Median [IQR] active cases in negative RT-PCR group and positive RT-PCR group were respectively 32.77 [27.23-34.89] and 608.77 [134.19-705.96] ($p < 0.001$).

We used 1413 (75%) cases to train the MLM for the prediction of the SARS-Cov-2 RT-PCR result. We used 471 (25%) cases to test the MLM for the prediction of RT-PCR result.

The new MLM was trained to predict SARS-Cov-2 RT-PCR positivity in terms of probability ranging from 0 (0%) to 1 (100%). When tested for the prediction of the SARS-Cov-2 RT-PCR, the MLM showed an accuracy of 95.75%, and the AURCC was 0.990 (95% confidence interval (CI)) [0.983 - 0.997] (Figure 2).

When tested for the prediction of the RT-PCR positive test, without using active cases as a feature, the MLM showed an accuracy of 80.0%, and the AURCC was significantly lower than the model trained with active cases 0.881 (95%CI) [0.852 - 0.911] ($p < 0.001$) (Figure 2).

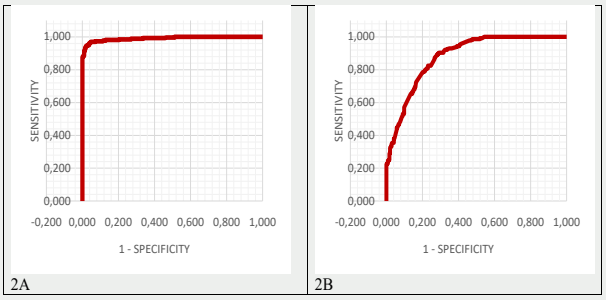

Figure 2. 2A: The area under the receiver operating characteristics curve of the machine-learning model tested for the prediction of a positive SARS-Cov-2 Reverse transcription polymerase chain reaction. **2B:** The area under the receiver operating characteristics curve of the machine-learning model tested for the prediction of the SARS-Cov-2 positive Reverse transcription polymerase chain reaction, excluding the geographic prevalence of Coronavirus Disease 2019 outbreak from the model.

The cross validation processed on the dataset showed a robust model as shown in Figure 4. The AURCC ranged from 0.958 to 0.984.

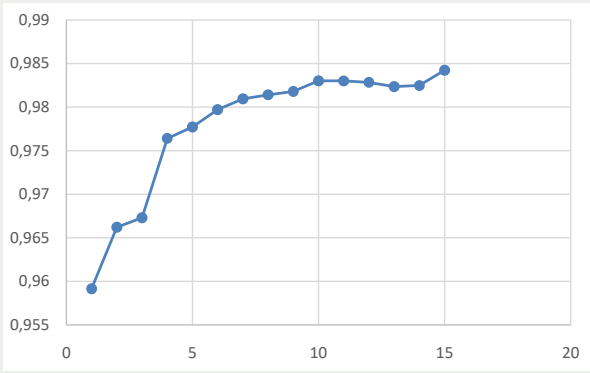

Figure 4. Cross validation process of the dataset using area under the receiver operating curve as metrics.

The weightings of the parameters showed a significant difference between the model using active cases versus the model trained without active cases (Figure 5).

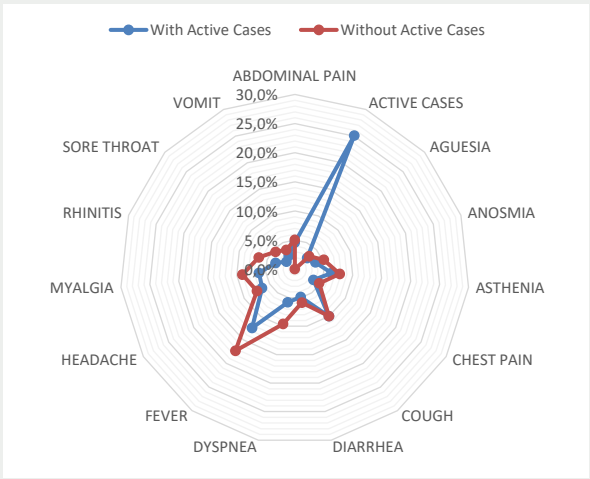

Figure 5. Features weightings regarding the two gradient boosting models, with active cases versus without active cases.

The three main features of the model with active cases were active cases incidence, fever and cough while the three main features in the model without active cases were fever, cough and dyspnea (Table 2).

Table 2. Detailed weightings of features used to train the gradient boosting models

	With Active Cases	Without Active Cases
ABDOMINAL PAIN	4.6%	5.0%
ACTIVE CASES	25.1%	0.0%
AGUESIA	2.9%	3.3%
ANOSMIA	3.7%	5.2%
ASTHENIA	6.4%	7.8%
CHEST PAIN	3.7%	4.8%
COUGH	10.0%	10.0%
DIARRHEA	4.9%	5.9%
DYSPNEA	5.8%	9.6%
FEVER	12.5%	17.3%
HEADACHE	6.5%	7.5%
MYALGIA	6.2%	9.0%
RHINITIS	3.5%	6.5%
SORE THROAT	1.9%	4.4%
VOMIT	2.5%	3.6%

The cut-off value of the new AI model for the prediction of RT-PCR positive test with the best Sensitivity + Specificity was 0.50 (50%).

The sensitivity of the new AI model for the prediction of RT-PCR positive test was 96.54% at the threshold value of 0.25 (25%). The specificity of the new AI model for the prediction of RT-PCR positive test was 99.05% at the threshold value of 0.75 (75%). For the threshold value of 0.5 (50%), the sensitivity of the new AI model for the prediction of RT-PCR positive test was 93.84% and the specificity 98.1%. Sensitivity, specificity, positive and negative predictive values and MCC for 0.25, 0.50, 0.75 thresholds of the AI model predictions are shown in Table 3. The MCC score at 0.916 indicates that the model is making strong and reliable predictions.

Table 3. Statistical evaluation of the machine-learning model for the prediction of positive Reverse Transcription Polymerase Chain Reaction test using three different thresholds.

Statistical parameter	25%[1]	50%[1]	75%[1]
Sensitivity	96.54%	93.84%	90.77%
Specificity	95.73%	98.10%	99.05%
Positive predictive value	96.54%	98.39%	99.16%
Negative predictive value	95.73%	92.85%	89.70%
Matthews coefficient of correlation	0.923	0.916	0.893

[1] Probability of positive SARS-Cov-2 RT-PCR predicted by the machine learning model.

Discussion

Our MLM based on clinical features and COVID-19 geographic prevalence was able to accurately predict positive RT-PCR of SARS-COV-2 infection. The high prediction accuracy of the MLM was supported by an AURCC of 0.990.

Since the first case confirmed in Wuhan (China) in December 2019, the onset of the new COVID-19 pandemic became a real challenge for the whole world healthcare systems [27]. The outbreak continued to spread all across the world and the WHO declared the pandemic as an international concern of public health emergency in the early 2020 [28]. The key strategy in terms of rapid control of the COVID-19 was the increased disease screening which led to early isolation and contact tracing [4]. With the 2020 COVID-19 outbreak, countless tests needed to be performed on symptomatic individuals, contacts and travelers [5]. The RT-PCR method suffered from shortcoming as the lack of accessibility in developing countries, or the significant delay to get the result, a time span that is not suitable regarding the effective public health containment measures [9]. The rapid tests, a serological method that detects anti SARS-COV-2 antibodies (Immunoglobulin (Ig) G/IgM) with a time to result of 20 minutes suffered from low sensitivity and drastically limited their usefulness to contain the pandemic outbreak [11]. To address this problem, machine-learning algorithms provide a realistic promising approach as a diagnosis and/or screening tool in the medical setting [16–21,29–31]. Recent studies showed that supervised and deep machine learning can be used as a reliable tool to support clinicians in the diagnosis step of the SARS-COV-2 infection, especially using CT scan images [32,33]. Using the deep learning CT scan images, some authors achieved a classification results for COVID-19 versus non COVID-19 of 0.996 AURCC (95%CI: 0.989 to 1.000) on Chinese control and infected patients, however, it is not clear whether authors included in the training set, CT scan images of viral pneumonia which can lead to confusion especially with COVID-19 images abnormalities [33]. In another study, deep MLM based on CT scan images achieved an AURCC of 0.99 and a recall (sensitivity) of 93%, however, the training dataset included only bacterial infection CT scan and therefore we can only assume that the MLM will be accurate even with other viral infections [34]. Since CT scan is not suitable for a screening strategy, some authors focused on clinical features to predict COVID-19 infection

[17,35]. A MLM based on healthcare workers clinical features achieved an AURCC of 0.754 (95%CI: 0.662 to 0.846), a recall (sensitivity) of 82.4% and a specificity of 59.2% [35]. Authors did not use contact with confirmed case, which could have increased the accuracy of the model, and limited their model on clinical features. Zoabi et al. [17] trained their MLM on clinical parameters but added contact with confirmed case to improve the accuracy of the model in the prediction of COVID-19 infection. They achieved an AURCC of 0.900 (95%CI: 0.892 to 0.905). When they trained and tested their MLM after filtering out features of high bias, they obtained an AURCC of 0.962, and contact with confirmed case had the most important impact on the model output [17]. Tse et al. [20] trained a MLM on a large number of features including age, sex, serum levels of neutrophil (continuous and ordinal), serum levels of leukocytes (continuous and ordinal), serum levels of lymphocytes (continuous and ordinal), result of CT scans, result of chest X-rays, reported symptoms (eg; diarrhea, fever, coughing, sore throat, nausea, and fatigue), body temperature, and underlying risk factors (eg; renal diseases, and diabetes mellitus). They also included a dataset of H1N1 influenza (viral respiratory disease transmitted by the type A Influenza virus family) patients in order to help the model distinguish COVID-19 patients from Influenza patients. The trained XGBoost model successfully distinguished COVID-19 patients from influenza patients with a recall (sensitivity) of 92.5% and a specificity of 97.9%.

In our study, we addressed the diagnosis support of SARS-COV-2 infection from another point of view. We aimed to focus on making a MLM that takes into consideration the incidence of COVID-19 active cases, in addition to clinical symptoms. This method raised a new concept that we called “dynamic disease screening”. The objective was to train the MLM to “dynamically” raise COVID-19 probability when the local incidence of active cases is high, and lower the COVID-19 probability when the local incidence of active cases is low, since the same symptoms may be related to other contagious diseases, which is instinctively a logical reflection that a physician would make. In the real time scenarios, the clinical symptoms will be provided by the user (patient or physician) and the previous day incidence data will be automatically collected by the application using the MLM model (or by the user regarding a communicated value of incidence in the own region when provided). In our model, the main features were COVID-19 active cases incidence, fever and cough. Fever and cough were the most frequent symptoms in the positive RT-PCR group. These results were confirmed by other studies showing fever and cough as the most common symptoms at early presentation [6,7]. We must acknowledge that these symptoms are common to several other diseases such as influenza or community acquired pneumonia. We think that is where the power of the incidence parameter comes into play in our model and puts this parameter at the top of the weighting. However, we cannot assert with certainty how the incidence parameter would help the model distinguish between COVID-19 and other clinically close diseases, since the model was not trained on a dataset including several diseases. The solution presented in this study is an early, accessible, simple and reliable tool to screen SARS-COV-2 infection using a supervised MLM. We trained the model to predict positive RT-PCR result as a gold standard for the diagnosis of COVID-19. The results of our study showed that our MLM was able to accurately predict positive RT-PCR of SARS-COV-2 infection. The high prediction accuracy of the MLM was supported by an AURCC of 0.990. As previously mentioned, several authors studied the potential applications of machine learning as a diagnosis support tool for COVID-19. The authors used different datasets, different features and different algorithms to train their models, therefore, it is very difficult to make a fair comparison regarding the performances. To make an objective and scientific comparison, it will be necessary to test the different machine learning models on the same population. These results are summarized in Table 4.

Table 4. Performances of several machine learning models for the COVID-19 screening published in literature

Authors	Country	Data sources	Number of cases	ML algorithms	Performances
Elhechmi et al.	Tunisia	Pasteur Institute, Tunis	1884	XGBoost*	AURCC [95%CI]: 0.990 [0.983-0.997] Sensitivity: 0.94 Specificity: 0.98 Accuracy: 0.96
Zoabi et al. [17]	Occupied Palestine	Occupied Palestine Ministry of Health	51831	Gradient boosting	AURCC [95%CI]: 0.900 [0.892-0.905]
Batista et al. [16]	Brazil	Brazil Ministry of Health	235	Neural network, RF*, LR*, SVM*, Gradient boosting	AURCC: 0.85. Sensitivity: 0.68. Specificity: 0.85; Brier Score: 0.16
Garcia et al. [18]	Brazil	Public Health Department of Florianópolis	3916	RF*	Accuracy [95%CI]: 0.66 [0.62-0.69] Sensitivity [95%CI]: 0.65 [0.57-0.75] Specificity [95%CI]: 0.66 [0.60-0.70]
Mei et al. [19]	China	18 medical centers in 13 provinces in China	419	CNN*, SVM*, RF*, MLP*	AURCC: 0.92
Tse Li et al. [20]	USA	Public data *	413	XGBoost	Sensitivity 0.925 Specificity 0.979
Avila et al. [21]	Brazil	Hospital Israelita Albert Einstein (HIAE - São Paulo, Brazil)	510	Naïve Bayes Classifier	Sensitivity: 0.767 Specificity 0.767

*Public data: https://github.com/yoshihiko1218/COVID-19ML, KNHIS: Korean National Health Insurance Service, SVM: Support Vector Machines, RF: Random Forest, LR: Logistic Regression, CNN: Convolutional Neural Networks, MLP: Multi-layer perceptron, AURCC: Area under the receiver operating characteristics curve, XGBoost: Extreme Gradient Boosting CI: Confidence Interval.

This study has strengths as the number of cases collected, physicians noticed the features, and therefore it is very unlikely that a present symptom was not reported. This study included data about COVID-19 daily active cases incidence per 100'000 inhabitants, and this parameter generated what we called "dynamic" screening process since the prediction probability of the disease generated by the MLM became variable depending on the geographic and time incidence of COVID-19 active cases. We must however acknowledge that this study has some limitations. We relied only on the data reported to the laboratory of virology of the Pasteur Institute of Tunis. These data have some missing features as contact with a COVID-19 case, time since onset of symptoms. This is a national study, which is subject to several specific parameters as the country policies toward COVID-19 pandemic or climate conditions, and we must be precautious before generalizing the MLM predictions to other countries. Finally, the high accuracy of our model may cause an overfitting issue when tested on independent cases and therefore this accuracy needs to be confirmed by other studies.

Conclusions

This study showed that a MLM was able to accurately predict SARS-COV-2 RT-PCR positivity using simple clinical and epidemiological features. Moreover, we introduced a new concept of "dynamic screening" since we showed a significantly improved sensitivity and specificity of the MLM when we included COVID-19 real time geographic incidence. All the features used in our model are commonly available and therefore can be used by citizens or physicians for a rapid, accessible and large screening of COVID-19 cases. In the context of world healthcare future pandemic emergencies, we believe that this MLM method can be very useful as a rapid and reliable screening tool for contagious diseases, especially in the developing countries.

Funding: This research received no external funding.
Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of Pasteur Institute of Tunis (protocol code 2020/27/I/LR16IPT, date of approval: 28th May 2021).
Informed Consent Statement: Informed consent was obtained by physicians in charge of patients at hospitals at the time of the nasopharyngeal swab for all subjects involved in the study.
Data Availability Statement: Due to the national law restrictions for the protection of personal data, the research database of this study cannot be made public. However, the research data may be sent for verification or for use in another study after a mandatory new validation by the Ethics committee of the Pasteur Institute of Tunis. Please send mail to the corresponding author.
Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785–94.
2. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975;405(2):442–51.
3. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One. 2017 Jun 2;12(6):e0177678.
4. Honein MA, Christie A, Rose DA, Brooks JT, Meaney-Delman D, Cohn A, et al. Summary of Guidance for Public Health Strategies to Address High Levels of Community Transmission of SARS-CoV-2 and Related Deaths, December 2020. MMWR Morb Mortal Wkly Rep. 2020 Dec 11;69(49):1860–7.
5. Recommendations for national SARS-CoV-2 testing strategies and diagnostic capacities [Internet]. [cited 2024 May 20]. Available from: https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-lab-testing-2021.1-eng
6. Pullen MF, Skipper CP, Hullsiek KH, Bangdiwala AS, Pastick KA, Okafor EC, et al. Symptoms of COVID-19 Outpatients in the United States. Open Forum Infectious Diseases. 2020 Jul 1;7(7):ofaa271.
7. Mehta OP, Bhandari P, Raut A, Kacimi SEO, Huy NT. Coronavirus Disease (COVID-19): Comprehensive Review of Clinical Presentation. Front Public Health [Internet]. 2021 Jan 15 [cited 2024 Oct 15];8. Available from: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.582932/full
8. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020 Feb 20;382(8):727–33.
9. Núñez I, Belaunzarán-Zamudio PF, Caro-Vega Y. Result Turnaround Time of RT-PCR for SARS-CoV-2 is the Main Cause of COVID-19 Diagnostic Delay: A Country-Wide Observational Study of Mexico and Colombia. Rev Invest Clin. 2022 Mar 15;74(2):071–80.
10. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020 Mar 17;323(11):1061–9.
11. Döhla M, Boesecke C, Schulte B, Diegmann C, Sib E, Richter E, et al. Rapid point-of-care testing for SARS-CoV-2 in a community screening setting shows low sensitivity. Public Health. 2020 May 1;182:170–2.
12. Zied Elhechmi Y. Medicine at the dawn of Artificial Intelligence. Tunis Med. 2022 May;100(5):354–5.
13. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of COVID-19. Plos one. 2020;15(6):e0235187.
14. Ahamad MdM, Aktar S, Rashed-Al-Mahfuz Md, Uddin S, Liò P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Systems with Applications. 2020 Dec 1;160:113661.
15. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals. 2020 Oct 1;139:110059.
16. de Moraes Batista AF, Miraglia JL, Donato THR, Chiavegatto Filho ADP. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. 2020;
17. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine. 2021;4(1):1–5.
18. Garcia LP, Goncalves AV, de Andrade MP, Pedebos LA, Vidor AC, Zaina R, et al. Estimating underdiagnosis of covid-19 with nowcasting and machine learning: Experience from brazil. medRxiv. 2020;
19. Mei X, Lee HC, Diao K yue, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature medicine. 2020;26(8):1224–8.
20. Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai JC, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC medical informatics and decision making. 2020;20(1):1–13.
21. Avila E, Kahmann A, Alho C, Dorn M. Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios. PeerJ. 2020;8:e9482.
22. Laboratory biosafety guidance related to coronavirus disease (COVID-19) [Internet]. [cited 2021 Jul 31]. Available from: https://www.who.int/publications-detail-redirect/laboratory-biosafety-

guidance-related-to-coronavirus-disease-(covid-19)

23. whoinhouseassays.pdf [Internet]. [cited 2021 Jul 31]. Available from: https://www.who.int/docs/default-source/coronaviruse/whoinhouseassays
24. Organization WH. Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases: interim guidance, 17 January 2020. World Health Organization; 2020. 7 p.
25. Epi InfoTM | CDC [Internet]. 2019 [cited 2021 Jul 31]. Available from: https://www.cdc.gov/epiinfo/index.html
26. Karim M, Rahman RM. Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. 2013;
27. Dzieciatkowski T, Szarpak L, Filipiak KJ, Jaguszewski M, Ladny JR, Smereka J. COVID-19 challenge for modern medicine. Cardiology Journal. 2020;27(2):175–83.
28. Sohrabi C, Alsafi Z, O'neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International journal of surgery. 2020;76:71–6.
29. Stokes K, Castaldo R, Federici C, Pagliara S, Maccaro A, Cappuccio F, et al. The use of artificial intelligence systems in diagnosis of pneumonia via signs and symptoms: A systematic review. Biomedical Signal Processing and Control. 2022 Feb 1;72:103325.
30. Mrabet S, Aloui K, Ben Jazia E. Development of a Web Application based on Machine Learning for screening esophageal varices in cirrhosis. Tunis Med. 2023;101(8–9):684–7.
31. Okiyama S, Fukuda M, Sode M, Takahashi W, Ikeda M, Kato H, et al. Examining the Use of an Artificial Intelligence Model to Diagnose Influenza: Development and Validation Study. Journal of Medical Internet Research. 2022 Dec 23;24(12):e38751.
32. Ozsahin I, Sekeroglu B, Musa MS, Mustapha MT, Uzun Ozsahin D. Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence. Comput Math Methods Med. 2020 Sep 26;2020:9756518.
33. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, et al. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. arXiv preprint arXiv:200305037. 2020;
34. Ying S, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. medRxiv. 2020 Feb 25;2020.02.23.20026930.
35. Tostmann A, Bradley J, Bousema T, Yiek WK, Holwerda M, Bleeker-Rovers C, et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. Eurosurveillance. 2020 Apr 23;25(16):2000508.