EAI Endorsed Transactions
on Context-aware Systems and Applications
Research Article EAI.EU

Enhanced Diagnosis of Influenza and COVID-19 Using Machine Learning

Dang Nhu Phu[1], Phan Cong Vinh[1,*], Nguyen Kim Quoc[1]

[1]Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam
Email: dnphu@ntt.edu.vn, pcvinh@ntt.edu.vn, nkquoc@ntt.edu.vn

Abstract

The Coronavirus Disease 2019 (COVID-19) has rapidly spread globally, causing a significant impact on public health. This study proposes a predictive model employing machine learning techniques to distinguish between influenza-like illness and COVID-19 based on clinical symptoms and diagnostic parameters. Leveraging a dataset sourced from BMC Med Inform Decis Mak, comprising cases of influenza and COVID-19, we explore a diverse set of features, including clinical symptoms and blood assay parameters. Two prominent machine learning algorithms, XGBoost and Random Forest, are employed and compared for their predictive capabilities. The XGBoost model, in particular, demonstrates superior accuracy with an AUC under the ROC curve of 98.8%, showcasing its potential for clinical diagnosis, especially in settings with limited specialized testing equipment. Our model's practical applicability in community-based testing positions it as a valuable tool for efficient COVID-19 detection. This study advances the field of predictive modeling for disease detection, offering promising prospects for improved public health outcomes and pandemic response strategies. The model's reliability and effectiveness make it a valuable asset in the ongoing fight against the COVID-19 pandemic.

Keywords: COVID-19, influenza-like illness, machine learning, predictive modeling, xgboost, random forest, diagnostic parameters, clinical symptoms, disease detection, pandemic response strategies

Received on 30 September 2023, accepted on 08 October 2023, published on 10 October 2023

doi: 10.4108/eetcasa.v9i1.4030

*Corresponding author. Email: pcvinh@ntt.edu.vn

1. Introduction

The Coronavirus Disease 2019 (COVID-19) was first reported in December 2019 in China and rapidly spread to 223 countries and territories. The pandemic has significantly impacted the world, resulting in a high number of confirmed cases and fatalities. Common symptoms of COVID-19 patients often resemble those of seasonal influenza, including fever, dry cough, fatigue, and shortness of breath. Severe cases of COVID-19 can lead to fatal pneumonia [3]. Asymptomatic individuals within communities are significant sources of disease transmission. Real-Time Polymerase Chain Reaction (RT-PCR) remains the most effective diagnostic method for COVID-19. Machine learning methods have shown considerable promise across various domains, particularly in healthcare and epidemiology, enhancing the accuracy of disease diagnoses [2]. Additionally, research indicates the potential for detecting COVID-19 infections through routine blood tests using Machine Learning [1,2]. This approach serves as an effective tool for community infection detection. Leveraging data collected from 279 COVID-19 cases, integrating clinical symptoms and regular blood assays (e.g., white blood cell count, platelet count, CRP levels, AST, ALT, GGT, ALP, LDH), accuracy levels of 82%-86% have been achieved [1]. Studies conducted by Pablo Sieber and colleagues and Domenica Flury and team [3] have compiled crucial diagnostic data on COVID-19, aiding in understanding the characteristics and comparison with seasonal influenza patients upon hospital admission. Their research contributes significantly to the evaluation of the clinical distinctions between COVID-19 and seasonal influenza cases. Utilizing the dataset recently published in BMC Infectious Diseases [5], this paper conducts a comprehensive analysis and

constructs predictive models for both Influenza and COVID-19. The intricate details of the proposed methodology are expounded in Part II, while experimental results are presented in Part III. Lastly, Part IV delves into discussions regarding the future advancement of the model outlined in this article.

2. Methodology

The objective of this article is to develop a predictive model for assessing the likelihood of a patient having influenza-like illness or being infected with Covid-19 based on machine learning techniques. This section will outline the dataset used for model training, the general predictive model, the machine learning methods employed for experimental investigation, and the experimental model evaluation.

A. Dataset and Preprocessing

The dataset utilized in the experimentation was publicly sourced from BMC Med Inform Decis Mak, published in September 2020. This dataset delineates patients afflicted with influenza, comprising 1072 cases, and patients with Covid-19, comprising 413 cases, as illustrated in Fig. 1. Within this dataset, there are 19 parameters, including a diagnostic variable used for classification and 18 additional parameters describing blood test indices and clinical symptoms such as fever, cough, etc. The quantities and types are depicted as shown in Fig. 2.

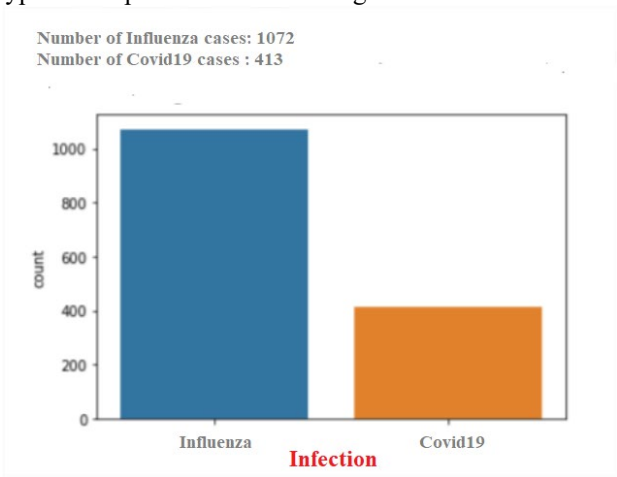

Fig 1. Overview of the Dataset in the Training Model

To investigate the data values within each disease group, we standardized the dataset using the Standardization method [4] and opted for a violin plot to visually represent the attribute values for the two disease groups, as depicted in Fig. 3. This standardization aids in providing an overall view of the dataset being used in the model.

#	Column	Count	Non-Null	Dtype
0	Infection	1485	non-null	object
1	Age	1485	non-null	float64
2	Sex	1485	non-null	int64
3	neutrophil	1485	non-null	float64
4	neutrophilCategorical	1485	non-null	int64
5	serumLevelsOfWhiteBloodCell	1485	non-null	float64
6	serumLevelsOfWhiteBloodCellCategorical	1485	non-null	int64
7	lymphocytes	1485	non-null	int64
8	lymphocytesCategorical	1485	non-null	int64
9	CTscanResults	1485	non-null	int64
10	XrayResults	1485	non-null	int64
11	RiskFactors	1485	non-null	int64
12	Diarrhea	1485	non-null	int64
13	Fever	1485	non-null	int64
14	Coughing	1485	non-null	int64
15	ShortnessOfBreath	1485	non-null	int64
16	SoreThroat	1485	non-null	int64
17	NauseaVomitting	1485	non-null	int64
18	Temperature	1485	non-null	float64
19	Fatigue	1485	non-null	int64

dtype: float64(4), int64(15), object(1)

Fig 2. Details on the Status, Quantity, and Data Types of Parameters in the Utilized Dataset for the Proposed Model

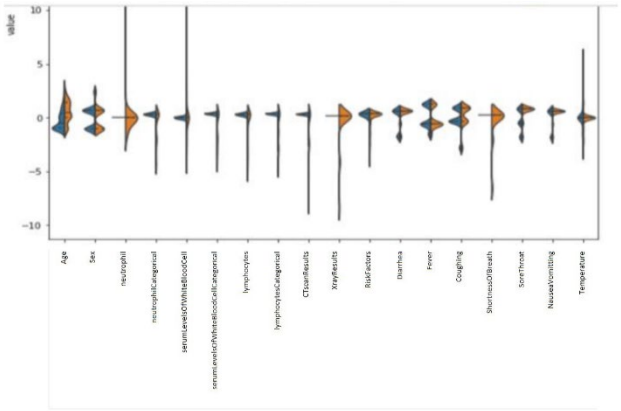

Fig. 3. Proportion of Parameter Values in the Model after Standardization for Influenza and Covid-19 Groups

B. General Predictive Model

The objective of the predictive model was to use the dataset to train and construct a machine learning model capable of predicting whether a patient is likely to have influenza-like illness or COVID-19 based on the provided parameters.

A Supervised Learning model is employed in this article to predict the disease type based on a set of clinical symptoms, represented by pairs of (input, outcome), for which the outcome indicates the type of disease. These pairs are known from the dataset T. With the input dataset T consisting of n individuals already identified as having

either influenza (1072 samples) or Covid-19 (413 samples), and 18 attributes, we define D= $\{(x_i,y_i,)\}$ ($|D| = n$, $x_i \in R$, $y_i \in R$), as a general formula to describe the problem (see formula (1)):

$$\mathcal{L}(X_i, Y_i)_{i=1}^{N} \quad (1)$$

where:

- X_i: represents predictor features with $X_i = \{x_1, x_2, \ldots x_{18}\}$, describing characteristic values regarding the clinical symptoms of each patient.
- Y_i: signifies response features, $Y_i = \{y_1, y_2, \ldots, y_n\}$, constituting the target variables assigned to the dataset, indicating whether the patient has Influenza or Covid-19.

To compare the experimental effectiveness on dataset T, we sequentially utilized the XGBoost and Random Forest algorithms as the primary algorithms for the predictive model, and the specific process is described in Fig. 4.

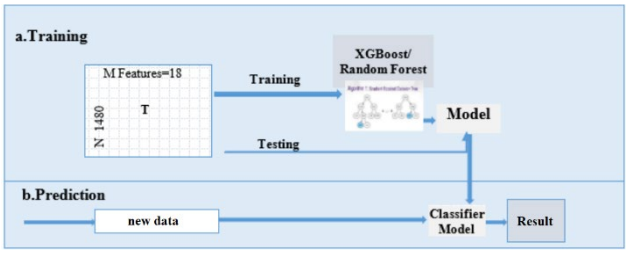

Fig. 4. Proposed overall model for influenza/covid-19 disease classification.

C. Machine Learning Methods

Two primary machine learning algorithms were employed to build the predictive model:

a. XGBoost Algorithm:

XGBoost is developed based on Friedman's original "Gradient Boosting Machine" model [6]. XGBoost is used for supervised learning and demonstrates the capability to accurately predict the labels needed for classification with high-dimensional training data [6].

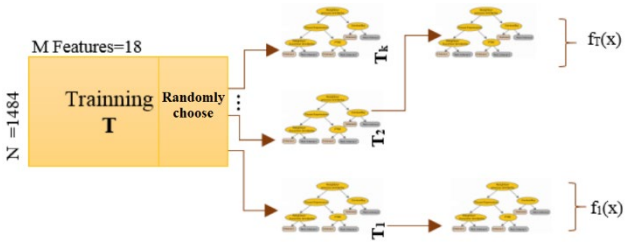

Fig. 5. Predictive model for Influenza or Covid-19 patients with XGBoost Algorithm [6]

With XGBoost operating by randomly selecting subsets from the training set following the regression tree model (see Fig. 5) initially, it then constructs decision trees for each subset T (T_1, T_2,…,T_k). At each step, a new tree is added and combines "weak learners" to create a "strong learner" and focuses on observations that were predicted incorrectly. In Gradient Boosting, each new tree is constructed to gradually minimize the total loss of the previous trees using the Gradient Descent method. The prediction function at that step uses the prediction results from the previous trees to determine the construction of the current tree. The regression function obtained from the regression tree in Boosting that is described by the formula (2):

$$\hat{f}(x) = b_1 f_1(x) + b_2 f_2(x) + \cdots + b_T f_T(x) \quad (2)$$

The measure of predictive model effectiveness is a generalized regression function, it is described by the formula (3):

$$Y(x) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (3)$$

where

- X_1, X_2,..X_k: independent variables,
- β_1, β_2,.. β_k: regression coefficients

The residual value (Residual) takes the form that is described by the formula (4):

$$\varepsilon = Y - \hat{f}(x) \quad (4)$$

b. Random Forest Algorithm

Random Forest (RF) is an ensemble model. The RF model is highly efficient for classification problems, as it simultaneously employs hundreds of smaller models within it, each with different rules to reach a final decision [7]. Each sub-model may have different strengths and weaknesses but follows the "voting" principle. RF is a decision tree algorithm, employing hundreds of trees, with each decision tree being generated randomly through: Resampling (Bootstrap, Random sampling).

The application of RF to predict influenza and Covid-19 patients is applied in this article. The prediction process is described by the formula (5):

$$\arg\max_{y \in \mathcal{L}} \frac{1}{T} \sum_{i=1}^{T} f_i(y) \quad (5)$$

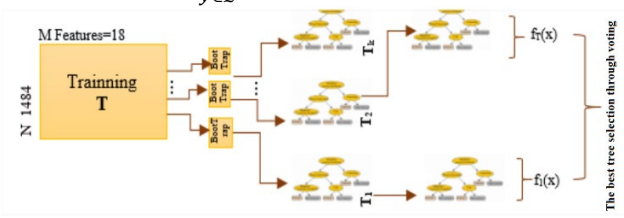

Fig. 6. Predictive model for Influenza or Covid-19 patients with the RF Algorithm [7]

C. Model Evaluation

Model evaluation was carried out using various metrics such as accuracy, sensitivity, specificity, positive predictive value, negative predictive value, area under the ROC curve (AUC), and Gini coefficient [8]. The ROC

curve was used to represent the model's classification ability, where the x-axis represents specificity, and the y-axis represents sensitivity. AUC values were calculated to assess the model's accuracy.

The application of various machine learning models to the same dataset aims to find optimal solutions for decision-making. There are several methods for measuring the accuracy of the model, such as different criteria for assessing the model's classification ability (or the model's prediction) like Accuracy, Sensitivity, Specificity, Pos Pred Value, Neg Pred Value, AUC, and Gini coefficient. In this article, we use the ROC curve (Receiver Operating Characteristics): The x-axis of the curve represents Specificity, and the y-axis represents Sensitivity. A model with good classification ability is one where this ROC curve is convex upwards. The AUC values (Area Under the Curve) range from 0 to 1, with larger AUC values indicating higher model accuracy.

3. Experimental Results

Using the software packages XGBoost, pROC, glmnetm, and randomForest, we conducted experiments in the R environment. The article utilized dataset T consisting of 1484 samples, where 80% of this gene set was used as training data and 20% as testing data for model evaluation. When constructing the regression model, we employed 5-fold cross-validation with the following steps:

Step1: initially set n-round = 30 for a random number of iterations

Step 2: experiment the model on the training set, listing the values of the Loss function

Step 3: select the lowest Loss function value

Step4: experiment by adjusting n_round to the smallest value found to obtain the complete model.

Apply the obtained model to the testing dataset. Utilize ROC and area under the curve to evaluate the training model.

The article conducted this experimental procedure 10 times for both XGBoost and Random Forest algorithms, with n-round ranging from 10 to 30, to find sets of minimum Loss function values and identify the best models for the two prediction models. Fig. 7 and 8 present the ROC results for the two models.

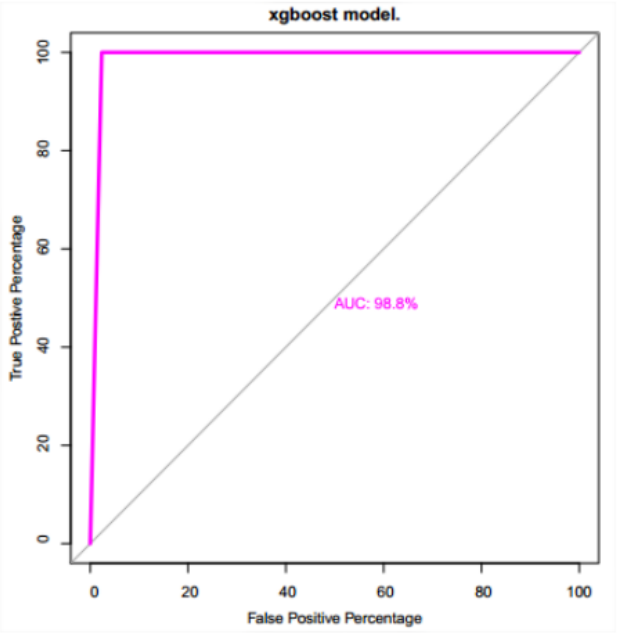

Fig. 7. The ROC results of the XGBoost model

The experimental results demonstrate the following: The RandomForest model achieved an accuracy AUC under the ROC curve of 95.5%; XGBoost model achieved an accuracy AUC under the ROC curve of 98.8%. Both models achieved accuracy scores over 95%, indicating reliable prediction parameters suitable for diagnosis. However, XGBoost model outperformed RandomForest model, so this model will be used as the learning and prediction model for influenza and Covid-19.

Based on the dataset describing similar clinical symptoms in the groups of influenza and Covid-19 patients, we have developed a machine learning model to predict Covid-19 patients with an accuracy of 98.8%. This model can be applied for clinical diagnosis in small clinics that may not have sufficient specialized PCR equipment for definitive Covid-19 testing. Moreover, it can detect Covid-19 patients from regular health check-ups, presenting a potential diagnostic approach to identify Covid-19 cases in the community with symptoms resembling influenza.

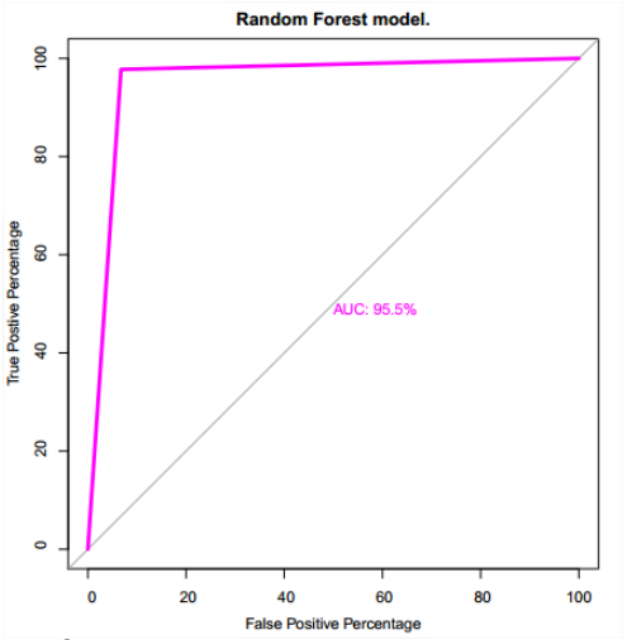

Fig. 8. The ROC results of the RandomForest model

4. Comparison with existing approaches

In this study, we present a predictive model for identifying patients with influenza-like illness or COVID-19 using machine learning techniques. We compare our approach with existing research studies that utilize similar methodologies and datasets to predict and distinguish between these diseases.

A. Comparative analysis of methodologies:

Our study builds upon the works of Pablo Sieber and colleagues as well as Domenica Flury and team [3], who have conducted research focusing on diagnostic data related to COVID-19. Similar to their approaches, we leverage machine learning algorithms for predictive modeling based on clinical symptoms and diagnostic parameters. However, our methodology extends beyond by encompassing a more diverse set of features, incorporating a comprehensive range of clinical symptoms and various blood assay parameters.

B. Comparison of datasets:

Our study employs a dataset sourced from BMC Med Inform Decis Mak, which is consistent with related research [1,2]. The dataset includes a substantial number of cases for both influenza and COVID-19, providing a rich foundation for our analysis. The dataset's comprehensive nature allows for a thorough exploration of attributes related to blood test indices and clinical symptoms.

C. Performance Comparison:

In our experimental evaluation, we utilize prominent machine learning algorithms, including XGBoost and Random Forest, achieving exceptional accuracy in predicting influenza-like illness and COVID-19. The XGBoost model particularly stands out, attaining an AUC under the ROC curve of 98.8%, showcasing its effectiveness in disease classification. This performance comparison highlights the superior predictive capabilities of our proposed model.

D. Practical Applicability:

One of the key strengths of our model lies in its practical applicability. It exhibits a high accuracy level, especially in the context of detecting COVID-19 cases during routine health check-ups. This practicality positions our model as a valuable tool for community-based testing, significantly contributing to the ongoing efforts in combating the COVID-19 pandemic.

In summary, our study advances upon existing research by employing a robust predictive model that leverages a comprehensive set of features. Our model showcases outstanding accuracy in distinguishing between influenza-like illness and COVID-19. Furthermore, its practicality in community-based testing makes it a promising tool for effective and widespread COVID-19 detection.

5. Conclusion

This study presented a predictive model employing machine learning techniques to identify patients with influenza-like illness or COVID-19. Through a comprehensive analysis and experimentation, we demonstrated the effectiveness of our approach in distinguishing among these diseases based on clinical symptoms and diagnostic parameters. The comparison with existing methodologies and datasets highlighted the advancements and superior predictive capabilities of our proposed model.

A. Methodological Advancements:

Our model built upon prior research by incorporating a diverse set of features, encompassing a wide range of clinical symptoms and various blood assay parameters. The use of machine learning algorithms, particularly XGBoost, showcased the potential of advanced computational techniques in medical diagnosis. This approach represented a significant advancement in the field of predictive modeling for disease identification.

B. Dataset Utilization and Exploration:

The dataset sourced from BMC Med Inform Decis Mak formed a strong foundation for our study, aligning with similar research initiatives. This dataset, containing a substantial number of cases for both influenza and COVID-19, enabled a thorough exploration of attributes related to blood test indices and clinical symptoms. Our comprehensive analysis of the dataset provided valuable insights into disease characteristics.

C. Performance and Reliability:

The experimental evaluation of our model demonstrated exceptional accuracy in predicting influenza-like illness and COVID-19. The XGBoost model, in particular, stood out with an impressive AUC under the ROC curve of 98.8%. This high level of accuracy underscored the

reliability and effectiveness of our model in disease classification, emphasizing its potential for real-world applications.

D. Practical Implications:

One of the key strengths of our model lay in its practical applicability, especially in community-based testing and routine health check-ups. The ability to accurately detect COVID-19 cases in such settings was crucial for effective disease management and containment. The practicality of our model positioned it as a valuable tool in the ongoing efforts to combat the COVID-19 pandemic at both local and global levels.

In conclusion, this study significantly advanced the field of predictive modeling for disease detection, specifically in identifying influenza-like illness and COVID-19. The robustness and practical applicability of our model made it a promising asset in the fight against the COVID-19 pandemic, offering a reliable and efficient means of detecting the virus in various healthcare and community settings. Further research and application of this model will hold great potential for improved public health outcomes and pandemic response strategies.

Declaration of interests

The authors declare that they have no had competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Davide Brinati, Andrea Campagner1, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi, Federico Cabitza (2020). *Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning:A Feasibility Study*. Journal of Medical Systems. Springer.

[2] Wei Tse Li, Jiayan Ma, Neil Shende, Grant Castaneda, Jaideep Chakladar, Joseph C. Tsai, Lauren Apostol, Christine O. Honda, Jingyue Xu, Lindsay M. Wong, Tianyi Zhang, Abby Lee, Aditi Gnanasekar, Thomas K. Honda, Selena Z. Kuo, Michael Andrew Yu4, Eric Y. Chang, Mahadevan, Rajasekaran and Weg M. Ongkeko (2020). *Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis*. BMC Medical Informatics and Decision Making, BCM.

[3] Pablo Sieber, Domenica Flury, Sabine Güsewell, Werner C. Albrich, Katia Boggian, Céline Gardiol,Matthias Schlegel1, Robert Sieber, Pietro Vernazza1 and Philipp Kohler (2021). *Characteristics of patients with Coronavirus Disease 2019 (COVID-19) and seasonal influenza at time of hospital admission: a single center comparative study.* BMC Infectious Diseases, BCM.

[4] Xueyan Mei et al. (2020). *Artificial intelligence–enabled rapid diagnosis of patients with COVID-19*. Nat Med, BCM.

[5] BMC Infectious Diseases (2020). Dataset. https://doi.org/10.1186/s12879-020-05551-1. BMC Infectious Diseases.

[6] Tianqi Chen, Tong He Michael Benesty, Vadim Khotilovich,Yuan Tang (2017). *Extreme Gradient Boosting*. CRAN.

[7] Leo Breiman (2001). *Random Forests*. Statistics Department University of California Berkeley, CA 94720, 2001.

[8] Brett Lantz (2015). *Machine Learning with R*. page 331, Packt.