Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/38/20230558

Clinical diagnosis of overlapping symptoms in COVID-19 based on machine learning model

Jingxuan Zhang

College of Arts and Sciences, New York University, New York, 10012, United States

jz4495@nyu.edu

Abstract. The global pandemic COVID-19 erupted and infected an estimated 10% of the world's population. Since vaccinations greatly reduced hospitalization rates, most countries removed the restrictive policies implemented to combat the virus. It has become a rather common illness with more than twelve thousand active hospitalizations. As a result, convenient COVID-19 diagnosis from diseases that display overlapping symptoms has become increasingly important. An effective method for patient self-diagnosis greatly reduces hospital presentation, saving time and medical resources. This study uses machine learning techniques to classify and predict several common respiratory diseases quickly and accurately. The author trains several machine learning models that attempt to predict four diseases based on their distinct clinical signs. An open-access database on Kaggle developed for this disease classification is selected and further processed via principal component analysis to decrease database dimension and pinpoint critical symptoms. Support Vector Machine Classifier (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF) models are used, and their performances are compared. Study results show that the LR model slightly outperforms the others. In conclusion, the effectiveness of the proposed method is proved for classifying the symptoms of patients with allergies, colds, flu, and Covid-19 in this study.

Keywords: Machine Learning, COVID-19, Overlapping Symptoms, Respiratory.

1. Introduction

It's been more than three years since the World Health Organization (WHO) declared COVID-19 as a pandemic. Despite its status as a global health emergency finally ending on May 11, 2023, due to a steadily decreasing number of active cases, coexisting with COVID-19 has become a reality due to its high infection rates. As a respiratory virus whose primary symptoms involve fever, cough, runny nose, etc., COVID-19 appears similar to many common illnesses such as colds, seasonal allergies, and the flu, each of which has its treatment methods. The most common COVID-19 screening methods are Reverse Transcription Polymerase Chain Reaction (RT-PCR) and antigen tests, both of which require the patient to present on-site, and they might have to wait for results due to processing delays or lab analysis. When symptoms are mild, patients can take over-the-counter medication for a quick recovery without presenting to the hospital, as long as they can correctly identify the type of illness. As such, demand for a convenient respiratory disease screening test is high.

Machine learning approaches were exploited to automate the screening process while maintaining accuracy. An important method of COVID-19 diagnosis bases itself on Computer Tomography (CT)

DOI: 10.54254/2755-2721/38/20230558

images of the lungs. A convolutional neural network (CNN) was employed to obtain satisfying predictive values based on the subtle lab-test differences between pneumonia and SARS-CoV-2 patients [1]. Another study uses the enhanced gray wolf optimization selection algorithm to increase feature selection accuracy before training a k-nearest neighbor classifier [2]. Asmaa H. Rabie et al. analyze chest X-ray (CXR) images using a Bayesian optimized CNN to obtain a radiation-free and cost-effective model [3]. Yang Xia et al. try fusing CXR with clinical features for rapid screening [4]. Random forest model, with the superior ability to exploit complex and nonlinear data, has also been applied to distinguish respiratory diseases [5]. Usually, these studies combine multiple test-based clinical data and laboratory results to increase model robustness and generalizability [6]. Machine learning has been documented to have the ability to accurately classify breath sounds [7]. Rumana et al. utilized frequency domain acoustic feature vectors [8]. To date, the most successful models have achieved a predictive performance of approximately 97% when evaluating COVID-19 alone. Two laboratory tests, monocyte count and basophil percentage, are the main differences between these two infections, but obtaining these results undermines the goal of easy classification [9]. Finally, Farrokh Alemi et al. explore this topic using Linear regression (LR) modeling. However, the dataset uses manually constructed external controls with neither matched baseline characteristics nor statistical analysis to assess data validity [10].

The main objective of this study is to classify with satisfying accuracy and false-positive rates the type of respiratory illness such as COVID-19, influenza, seasonal allergies, and common cold based on patients' readily available symptoms. Since additional requirements limit the selection of test-based characteristics, the models attempt to learn from various indirect, less related, and possibly confounding observable symptoms. Thus, selecting the correct features during the data preprocessing phase is crucial. Support Vector Machine (SVM), Naive Bayes (NB), LR, and Random Forest (RF) are introduced and evaluated. The experimental results show that the LR model is among the best in terms of prediction results and efficiency. In repeated runs with different random states, the SVM model can achieve similar accuracy scores to the LR model, but the overall performance cannot be matched. The loss function of the LR model had to be changed to support non-binary target variables, but unlike the other two models, it can handle binary independent variables efficiently and effectively. It was demonstrated that the LR model can classify the type of respiratory disease experienced by users with high accuracy based on self-assessed clinical symptoms.

2. Methodology

2.1. Dataset description and preprocessing

This paper uses Kaggle's open-access dataset that records 44,452 patients diagnosed with allergies, cold, flu, and COVID-19 [11]. Filtered from the dataset contains twenty different symptoms, all of which may be assessed without requiring professional equipment and assistance. The traits are presented below in Table 1. The absence or presence of a symptom is represented as 0 and 1 respectively. The target parameter, denoted "type," contains four possible values: allergy, cold, COVID-19, and flu. These are converted into numerical values suitable for machine learning recognition. A principal component analysis (PCA) is done on the dataset to rule out features that explain less than 5% of the total variations, leaving nine relatively more significant characteristics. Later, adjusting the strictness of this filter allows for model tuning and model robustness enhancement.

Table 1. The traits of the data.

COUGH	MUSCLE_ACHES	TIREDNESS	SORE_THROAT
STUFFY_NOSE	FEVER	NAUSEA	VOMITING
SHORTNESS_OF_BREATH	DIFFICULTY_BREATHING	LOSS_OF_TASTE	LOSS_OF_SMELL
ITCHY_EYES	ITCHY_MOUTH	ITCHY_INNER_EAR	SNEEZING
RUNNY_NOSE	DIARRHEA	ITCHY_NOSE	PINK_EYE

DOI: 10.54254/2755-2721/38/20230558

2.2. Proposed approach

SVM, NB, LR, and RF are common machine-learning algorithms used in classification tasks. Specifically in the context of this dataset that consists of all binary variables, the LR model stands out as the algorithmically most intuitive and effective model. On the other hand, it is especially prone to overfitting training data due to the high linear dependence of given variables. Thus, all four machine learning algorithms are trained and tested based on the same dataset. Figure 1 below contains a step-by-step process through which the dataset is processed, trained, and tested. Before the dataset is randomly split into two for training and testing, the traits first go through a principal component analysis of the covariance matrix to leave out traits that explain less than 2% of data variations. Actively reducing the dimension helps exclude unnecessary variables and speed up the learning process. Finally, the author uses confusion matrices and mean cross-validation scores to compare model performances.

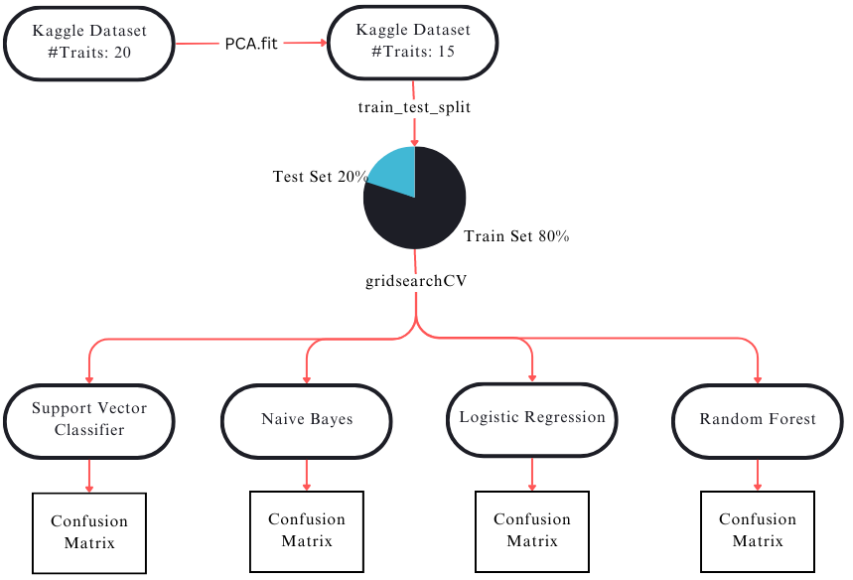

Figure 1. The pipeline of the study.

2.2.1. SVM. SVM is a supervised machine learning technique that effectively classifies examples into distinct categories. When training the model, the classifier treats each entry as a vector of multiple dimensions. Then, it achieves data point separation by maximizing the distances between points of distinct categories. When manually selecting a linear kernel, the model selects two parallel hyperplanes at the boundary of the data points and calculates the distances between them. Through continuous computation, the computer determines the maximum distance between the hyperplanes. As such, new data of similar traits tend to fall within the area between the hyperplanes of respective categories. For this multi-class classification task, the one-versus-all strategy is used. The training dataset consists of randomly selected samples from different categories, allowing the model to effectively separate them. Finally, the SVM classifier employed in this study chooses a linear kernel.

2.2.2. NB. NB algorithm bases itself on Bayes' theorem and assumes all features are independent. This assumption greatly reduces computation complexity, losing some strength in exchange for efficiency. The Bernoulli NB (BNB) variant is specially designed for binary feature data. To train a BNB model, it calculates the probabilities for each target class by estimating the prior probabilities and the likelihood probabilities of the features given in each category. BNB classifies new data by calculating the posterior probabilities for each kind of illness based on input features. Then, it outputs the one with the highest conditional probability. This simple algorithm is highly automated with little room for tuning. However,

DOI: 10.54254/2755-2721/38/20230558

its high learning speed and sensitivity to input data indicate whether the best variables are selected and a nice baseline comparison with other models.

This paper uses the extended multinomial logistic regression model to predict the probabilities. The model's weights or coefficients are optimized to maximize the likelihood of observing the training data given the model's predictions. Like binary logistic regression, multiclass logistic regression can be trained using maximum likelihood estimation. The one-vs-rest (OvR) and the SoftMax (or multinomial) approach. The latter estimates the probability of the outcome belonging to each class simultaneously using the SoftMax function, which normalizes the output probabilities across all classes. The SoftMax approach is chosen given its superior performance in dealing with high-dimensional data.

2.2.3. RF. The RF approach combines the predictions of multiple decision trees to produce classification results. In this case, individual decision trees in the random forest make their predictions for the diagnosis outcome of a patient, considering the 20 features. Once all the decision trees have made their predictions, the final output is determined by selecting the most frequently occurring diagnosis among the individual tree predictions. This majority voting technique requires an appropriate number of trees, but overfitting concerns rise along with increasing accuracy. Adjusting for the maximum number of trees and the minimum sample size in each leaf helps optimize the training result. Finally, the model has a total of 68 estimators, meaning it consists of 68 individual decision trees. The maximum depth of each decision tree, which controls how deep the tree can grow, is set to 5. The optimal minimum sample size is 17. These parameters help control the complexity and generalization ability of the random forest model as a whole.

3. Result and discussion

By adjusting the parameters, each model has close to 90% accuracy, except for NB. The experimental result of each model can be found in Table 2. Overall, LR has the highest accuracy among the three machine-learning models.

Table 2. The experimental result of each model.

Model	SVM	NB	LR	RF
Best Score	0.9314	0.6313	0.9320	0.9280
Weighted Precision	0.89	0.71	0.93	0.89
Weighted Recall	0.94	0.62	0.94	0.93
Training time	5m 55.7s	4.6s	28.6s	6m 33.2s

As expected, NB’s independence assumption has saved huge amounts of calculations. These models run a similar number of trials, but their training times differ by order of magnitude. However, the challenge of this classification task originates from the close proximity of clinical symptoms among respiratory diseases. The independence assumption thus severely harms model effectiveness. The SVM model performs decently, although the training speed reflects the model’s limitations when run on high dimensional binary data. Since SVM maps each entry to a vector, binary data ignores all values between 0 and 1 and wastes the corresponding dimension, leading to unnecessary training complexity. The training time for the RF classifier differs drastically depending on the range of parameters tested. When the maximum depths of each tree and the maximum number of trees increase, training complexity rises quickly. Because of this, having a good estimate of the suitable tree sizes beforehand is crucial to ensuring efficiency. For this dataset of size 20*44,452, the training takes more than 160 minutes if all reasonable tree sizes are considered. On the other hand, any combination of parameters greater than the optimal pair causes overfitting problems that drastically affect model precision.

Finally, the LR model ranks high in both prediction results and efficiency. In repeated runs with different random states, the SVM model could reach similar precision scores as the LR model, but the overall performance cannot be matched. The LR model’s loss function has to be altered to support a

DOI: 10.54254/2755-2721/38/20230558

non-binary target variable, but it works with binary independent variables efficiently and effectively, unlike the other two models. The result proves that the LR model can classify for its users the type of respiratory disease they experience based on self-assessed clinical symptoms with high accuracy. When additional data is provided, the model can easily be retrained to suit patients of corresponding regions and backgrounds.

4. Conclusion

The author compares four machine learning classifiers to output the most probable respiratory disease given clinical symptoms. After adjusting for parameters to improve learning curves and reduce overfitting, the LR model stands out with a slightly better overall performance regarding precision and recall. Its cross-validation mean score is 0.60% higher than that of the RF model. First, certain clinical traits that the models focus on, such as fever, shortness of breath, pink eyes, etc., are signs of immune responses caused by illnesses unrelated to the respiratory system. Expanding the scope of our predictive model increases the complexity as well as possible errors induced by inaccurate assessment of patient symptoms, which is more useful in real-world scenarios where users of this model cannot assume what type of disease they expect. Another viable improvement is that the well-performed models can be ensembled to achieve better overall accuracy. In exchange, model interpretability is affected when multiple machine-learning techniques are sequentially applied. The author plans to overcome the issue and introduce this technique in the future to generate more accurate classification results.

References

[1] Zhao W Jiang W Qiu X 2021 Deep learning for COVID-19 detection based on CT images Sci Rep 11s; p 14353
[2] Rabie A H Mohamed A M Abo-Elsoud M A Saleh A I 2023 A new Covid-19 diagnosis strategy using a modified KNN classifier Neural Comput Appl 2: pp 1-25
[3] Aslan M F Sabanci K Durdu A Unlersen M F 2022 COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization Comput Biol Med 142: p 105244
[4] Xia Y Chen W Ren H et. al. 2021 A rapid screening classifier for diagnosing COVID-19 Int J Biol Sci 17(2): pp 539-548
[5] Guo X Li Y Li H et. al. An improved multivariate model that distinguishes COVID-19 from seasonal flu and other respiratory diseases Aging (Albany NY) 12(20): pp 19938-19944
[6] Zhou X Wang Z Li S 2021 Machine Learning-Based Decision Model to Distinguish Between COVID-19 and Influenza: A Retrospective, Two-Centered, Diagnostic Study Risk Manag Healthc Policy 14: pp 595-604
[7] Belkacem Abdelkader Nasreddine 2021 End-to-End AI-Based Point-of-Care Diagnosis System for Classifying Respiratory Illnesses and Early Detection of COVID-19: A Theoretical Framework Frontiers in Medicine 8
[8] Islam R Abdel-Raheem E Tarique M 2022 A study of using cough sounds and deep neural networks for the early detection of Covid-19 Biomed Eng Adv 3: p 100025
[9] Chen J Pan Y Li G et. al. 2021 Distinguishing between COVID-19 and influenza during the early stages by measurement of peripheral blood parameters J Med Virol 93(2): pp 1029-1037
[10] Alemi F Vang J Wojtusiak J et al. 2022 Differential diagnosis of COVID-19 and influenza PLOS Global Public Health 2(7): p e0000221
[11] Walter Conway0 2021 COVID, FLU, COLD Symptoms. Kaggle. https://www.kaggle.com/datasets/walterconway/covid-flu-cold-symptoms