기계학습기반의 코로나 진단 및 증상 분석

김예담, Stuart Trivino
ykim2022@chadwickschool.org, strivino@chadwickschool.org

Machine Learning based COVID-19 Diagnosis and Symptom Analysis

Yedam Kim, Stuart Trivino
Chadwick International School

Abstract

The recent COVID-19 pandemic has accentuated the need for faster and more accurate ways of diagnosing certain diseases for there to be safer and more effective early responses that help to prevent a total outbreak. In this work, we would like to approach this issue through machine learning algorithms to investigate whether or not they could serve as a viable replacement for conventional diagnosis. Through a process of training and testing various algorithms, we analyzed how successfully they can predict a patient's COVID-19 diagnosis based on a list of symptoms and also identified which algorithm is the most effective at doing so. If the necessary data, containing the symptoms and diagnoses of different cases, is provided, this method can be utilized to make a probable diagnosis of any disease besides COVID-19. This method can be used in conjunction with or in lieu of conventional diagnosis depending on the situation: if there is a lack of testing facilities or test kits, this method can be employed as it is inexhaustible and it could also be used in situations where a conventional diagnosis is proven to be inaccurate.

1. Introduction

SARS-CoV-2, better known as COVID-19 or the Coronavirus, is an infectious respiratory illness that was first discovered in Wuhan, China back in December 2019. It has infected approximately 299 million people worldwide and killed four and a half million people as of now in October 2021. Symptoms of this disease include fever, fatigue, dry cough, dizziness, loss of smell or taste, joint pain, chest pain, and more. Long-term effects of this disease may include damages to organs such as the lungs which can ultimately lead to death in some case. Global efforts have been made to successfully diagnose and treat this disease, but the challenge has yet to be fully overcome.

Machine learning is a type of AI(Artificial Intelligence) that uses statistical analysis to recognize patterns in massive amounts of data to mimic human decision-making. There are various algorithms that are widely used by individuals and companies for a myriad of purposes ranging from targeting advertisements to detecting spam[1]. In this work, we apply classification algorithms for to predict COVID-19 infection. It should be assumed that diseases similar to COVID-19 will arise again in the future. In preparation for events such as this, machine learning can be used in situations where testing is inaccurate or impossible to effectively diagnose the disease. This would help control the spread of the disease and administer appropriate treatments more efficiently, leading to further conservation of life.

The work begins with acquiring an appropriate data set that will be used to train and test the different algorithms. After being processed, the data will be visualized to gain a general understanding of the relationship between the symptoms and the diagnosis. For our work, we explore five different classification algorithms for training and validation. After optimizing and testing the models, we analyze the performance through a matrix and four different metrics. Finally, a rational conclusion will be reached, and possible extensions to this work will be suggested.

2. Machine Learning - Classification

Classification problems are a form of supervised learning that are used for identifying the relationship between the data and its category. The most common algorithms used for this problem are KNN(K-Nearest Neighbors), decision tree, random forest, Naïve Bayes, and SVM(Support Vector Machine).

KNN is an algorithm that makes its predictions by referencing cases that are similar to the one in question. The outcomes of the neighboring cases are referred to and the majority of their outcomes are used to make a prediction[2].

Decision tree makes its prediction based on a series of questions that are structured in a flowchart-like manner. The questions are formed through training and a case would be processed through the list of questions, and based on the

answers, a prediction would be made[2].

Random forest utilizes multiple smaller decision trees to make its predictions. Each decision tree produces a prediction, and then the majority is chosen as the final prediction[3].

The Naïve Bayes makes its decision based on probability. It makes use of the Bayes' Theorem which calculates the probability of a result, A, given that a certain feature, B, is present[2]. This probability is then used to make a prediction.

SVM utilizes what's called a decision boundary, which is a line drawn between the cases in reference to their outcomes. A case would be placed on a plane and the outcome is predicted based on rather it is above or under the decision boundary [3].

3. COVID-19 Symptom Analysis

3.1 Data Collection

The data set for this work needs to contain different cases of COVID-19 diagnoses where the patient either tested positive or negative for COVID-19, accompanied by the list of symptoms that the patient had or didn't have. It also needs to contain a sufficient number of cases to both train and test the algorithms.

A data set that met these specifications was acquired from the data collected by the Indian Ministry of Health and Family welfare about the testing results of suspected COVID-19 patients in May 2020[4]. There are twenty-one different features within the data set and 5434 different cases of patients that either tested positive or negative for COVID-19. Out of the twenty-one different features, only eight are direct symptoms of COVID-19: breathing problems, fever, dry cough, sore throat, running nose, headache, fatigue, and gastrointestinal (symptoms such as vomiting and diarrhea). These features, and the one regarding the COVID-19 test results, are the ones that will be used within the data set.

3.2 Feature Engineering

Feature engineering, also called data preprocessing, refers to the process of manipulating the data to transform it into a focused and suitable format for an algorithm. This is a simple yet crucial step in creating a successful model.

① Relabeling of all the different features within the data set (Figure 1)

② Removal of the unnecessary features to solely identify the relationships between the symptoms and COVID-19

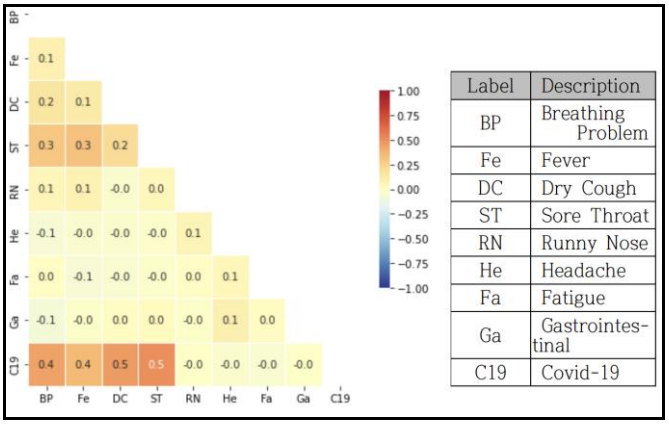
③ Binarization of the data to process the data more efficiently and effectively through a numerical representation

④ Division of the data into two sets: train set and test set

3.3 Visualization

The data is visualized to gain a general understanding of the correlations between the symptoms and the diagnoses before any predictions are drawn. The relationships will be visualized through two different types of graphs: heat map and bar graph.

Heat maps are a method of visualization that shows the magnitude of a phenomenon through colors[5]. In this case, the phenomena are the degrees of correlation between each symptom in relation to one another and COVID-19. The spectrum of colors between red and blue illustrates how strong the correlation is between the features: the redder color indicates a stronger correlation while the bluer color indicates a weaker correlation. In addition to the colors, the correlation coefficient is also shown on a scale of positive one to negative one.

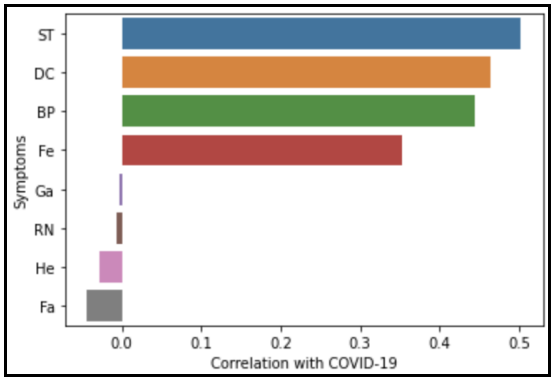


Label	Description
BP	Breathing Problem
Fe	Fever
DC	Dry Cough
ST	Sore Throat
RN	Runny Nose
He	Headache
Fa	Fatigue
Ga	Gastrointes-tinal
C19	Covid-19

(Figure 1) Correlation between Symptoms and COVID-19

The strongest correlation appears to be between the symptoms such as dry cough, sore throat, fever, and breathing problems in relation to COVID-19. Aside from that, there are no other notable correlations besides the ones between sore throat in relation to breathing problems and fever. All other correlations are either minuscule or zero, indicating that there isn't much correlation between most symptoms in relation to one another and COVID-19.

The purpose of the bar graph is to identify the correlations of each symptom solely in relation to COVID-19. This will provide a more detailed insight into the degrees of correlation that were previously generalized through the heat map.



(Figure 2) Correlation of Symptoms with COVID-19

Sore throat shows the highest degree of correlation with COVID-19 followed by dry cough, breathing problems, and

fever. These are the symptoms that are the most common with cases where the patient tested positive for COVID-19 and thus will be given the most amount of weight by the algorithms when making their predictions.

The symptoms that show a negative correlation will be taken into account by the algorithms in predicting that the patient might not have COVID-19. This can be interpreted as the limitation of the data set as a presence of a symptom cannot indicate the absence of a disease, signifying that more data is required to discern the absolute relationships between the symptoms and COVID-19.

4. Training & Testing for Diagnosis

4.1. Training and Validating

This is the main aspect of the work where five different machine learning algorithms are trained. The algorithms are KNN, decision tree, random forest, Naïve Bayes, and SVM. These five algorithms will be trained identically using the train data set and then validated using the K-Fold Cross Validation method to identify any errors to be fine-tuned before being tested.

Validating each of the algorithms after their initial training is a necessary step to take before testing. The validation scores indicate if there are any major issues that must be fixed and are also referred to when optimizing the algorithms. The method used is called the K-Fold Cross Validation which involves splitting the train set K times and validating the algorithms with all the divided sets. Then the validation scores from each of the sets are averaged to produce an average validation score for the algorithm. This is mainly done to avoid bias during the validation process.

<Table 1> Average cross validation scores

	KNN	Decision tree	Random forest	Naive Bayes	SVM
Average validation score(%)	97.18	98.10	97.96	89.26	97.27

All the average validation scores of the algorithms indicate that there aren't any major issues that have to be addressed as the scores approximately range between 90 and 100%, which is moderately high. This also indicates that there won't be a lot of room for optimization as the algorithms are already yielding good results.

There is a default setting of all the hyperparameters for most algorithms; however, they can and should be fine-tuned to achieve the best results. To fine-tune these hyper-parameters, a module called GridSearchCV[7] will be used to find the most optimal ones for each algorithm except Naïve Bayes because the Naïve Bayes utilizes probability calculated from an equation to predict the outcomes rather than trained parameters.

The validation scores acquired from the optimization indicate that the default hyperparameters yield the best results for all four algorithms. This means that the default hyperparameters are already the most well-suited for this task, and thus will be the ones used in testing the algorithms.

4.2. Testing and Result Analysis

The test data set will be inputted into the models, and they will produce their predictions based on their training and in accordance with their default hyperparameters. These predictions will be recorded to later be used for analysis. The results of the testing will not be directly observed but rather assessed through a method involving a matrix and various metrics.

The performances of the algorithms will be quantified through four distinct metrics: accuracy, precision, recall, and F1. These metrics are numerical values that are calculated with different equations that are drawn from what's called a confusion matrix.

The confusion matrix is a two-by-two matrix that compares the actual target values with the ones predicted by the algorithms[6]. It's divided into four quarters, each representing a possible outcome: true positive(TP), true negative(TN), false positive(FP), and false negative(FN). TP and TN are when the algorithms accurately predict a patient to have or not have COVID-19, while FP and FN are when it inaccurately predicts so. FP is also called a type one error and FN is called a type two error[6].

<Table 2> Confusion matrix results

	KNN	Decision tree	Random forest	Naive Bayes	SVM
True Positive	175	169	197	159	179
True Negative	866	857	858	823	866
False Positive	31	33	9	47	27
False Negative	4	17	12	47	4

The random forest has the most TP predictions with 197 cases while the KNN and SVM are tied with having the most TN predictions at 866 cases. On the other hand, the Naïve Bayes has the most predictions for both FP and FN, which indicates that it possesses the most type one and type two errors out of the five algorithms.

<Table 3> Metrics scores

	Score(%)			
	Accuracy	Precision	Recall	F1
KNN	96.75	84.95	97.77	90.91
Decision tree	95.35	83.66	90.86	87.11
Random forest	98.05	95.63	94.26	94.94
Naive Bayes	91.26	77.18	77.18	77.18
SVM	97.12	86.89	97.81	92.03

The accuracy score is the most important out of the four metrics as it's the most representative of the algorithm's performance. It reflects how often the algorithm is making a correct prediction. The random forest has the highest accuracy score with 98.05% while the Naïve Bayes has the lowest with 91.26%. This signifies that the random forest makes correct predictions the most often while the Naïve Bayes makes them the least often.

Precision score indicates how many of the predicted positive cases are actually positive. This score is significant for situations when the FP is a higher concern than FN. Similar to the accuracy score, the random forest has the highest precision score with 95.63% while the Naïve Bayes has the lowest with 77.18%. This means that out of the all

predicted positive cases, the random forest had the greatest number of actual positive cases while the Naïve Bayes had the least.

Recall score indicates how many of the actual positive cases were correctly predicted by the algorithms. It's opposite to the precision score in that it's mainly referenced when the FN is a higher concern than FP. For the recall score, the SVM has the highest score with 97.81%, and the Naïve Bayes has the lowest score with 77.18%. This signifies that the SVM correctly predicted the greatest number of actual positive cases while the Naïve Bayes predicted the least.

F1 score is the harmonic mean between the precision and recall metrics. It helps to gauge both metrics simultaneously through a single score. For the F1 score, once again, the random forest has the highest score with 94.94%, and the Naïve Bayes has the lowest with 77.18%. This means that overall, the random forest has the best balance of prediction and recall while the Naïve Bayes has the worst.

4.3. Discussion

Looking at the results, the most superior algorithm seems to be apparent as the random forest achieved the best scores for the accuracy, precision, and F1 metrics. However, it cannot be stated that the random forest is the best algorithm for this task as the recall metric should also be taken into account considering the type two error; in this context, a type two error is very critical as it would mean that the patient would be falsely informed that they don't have COVID-19 when they actually do, possibly leading to further spread of the virus. With this in mind, the algorithm that has scored the highest for the recall metric is the SVM, which has also ranked second in all the other metrics.

One other notable detail is that the Naïve Bayes has a noticeably inferior performance compared to the other four algorithms. It can be assumed that this is because of the mechanism behind the algorithm; when the Naïve Bayes calculates a probability using the Bayes' Theorem, all the features are given equal weighting. This mechanism isn't adaptable to the degrees of correlation between the features, which could possibly be the reason for these results.

The degree of success of most AI algorithms largely depends on the quality and quantity of the data provided. The larger the data set, the more the algorithm can learn and perfect its parameters through experience. The data set provided for the algorithms in this work can be considered limited as shown by the negative correlation between some of the symptoms and COVID-19. The number of cases, 5434, is relatively small compared to the size of the data sets used for other AI programs. The data set is also outdated as the list of symptoms is in reference to an old list published by the World Health Organization. They have since updated their list to include new common symptoms such as loss of taste or smell, which are not included in this data set.

5. Conclusion

The purpose of this work was to explore whether or not machine learning can be used to accurately diagnose COVID-19 based on a patient's symptoms. Through an intricate process of machine learning modeling and performance analysis, we've found that machine learning can be used to make accurate COVID-19 diagnoses purely based on a list of symptoms.

One way to expand upon this investigation is to design a model that predicts the mortality risk of a COVID-19 patient based on factors such as age, gender, symptoms, and pre-existing health conditions. This would require a very extensive data set that includes all the previously mentioned information, but the overall process and methodology would be very similar to the ones used in this work. Theoretically, this program would be applied after a diagnosis is given and would help medical staffs in assessing the urgency of the patient, leading to a more appropriate and faster treatment.

References

[1] G. Meyer, G. Adomavicius, P. E. Johnson, M. Elidrisi, W. A. Rush, J. M. Sperl-Hillen, P. J. O'Connor, "A Machine Learning Approach to Improving Dynamic Decision Making," Information Systems Research, Vol.25, No.2, 2014.

[2] A. Singh, M. N. Halgamuge, R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," IJACSA, Vol.8, No.12, 2017.

[3] P. Singh, S. P. Singh, D. S. Singh, "An Introduction and Review on Machine Learning Applications in Medicine and Healthcare," IEEE Conference on Information and Communication Technology, 2019.

[4] https://www.kaggle.com/datasets

[5] J. Zenko, "When (and Why) to Use Heat Maps," Premium Reporting & Data Analytics - Dundas Data Visualization, 22 Aug. 2018, https://www.dundas.com/resources/blogs/best-practices/when-and-why-to-use-heat-maps.

[6] A. Bhandari, "Confusion Matrix for Machine Learning," Analytics Vidhya, 17 Apr. 2020, www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/.

[7] M. Sharma, "Grid Search for Hyperparameter Tuning," Medium, Towards Data Science, 21 Mar. 2020, https://towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec.