UC-36 Using Machine Learning Techniques to Predict RT-PCR Results for COVID-19 Patients

Abstract

COVID-19 continues to prove to be unstoppable, even over a year into the pandemic. Using RT-PCR patient data from the early onset, we set out to improve the accuracy of the initial results using WEKA Machine Learning.

Introduction

In December of 2019, the world was beginning to see the rise of a new variation of coronavirus (SARS-CoV-2) emerging in Wuhan, China [1]. The virus, also known as COVID-19, is characterized by its ability to attack the lungs. The danger with this virus is the way that it effects individuals differently. Most cases involve mild symptoms, but there are a few high-risk groups, such as the elderly and those with compromised immune systems, where the virus is deadly. Towards the end of 2019, COVID-19 spread to 220 countries which prompted the calling of a global pandemic. Symptoms include fever, cough, shortness of breath, sore throat, and/or new loss of taste or smell [2]. Although most patients will only have mild symptoms, COVID-19 has the ability to cause pneumonia and complete respiratory failure.

In the earlier days of the virus, scholars began to explore different methods of applying machine learning to this problem, leading to a surge of research in the subject. Machine Learning has been involved in experiments with the prediction of test results, using both image classification and data classification, tracking the spread of the virus to improve hospital readiness, and even applications of finding a cure for the virus. In present times, the virus is still the hot topic and better diagnostic systems are still needed to help the domain of medicine.

According to [3], as of February 2021, there have been 108 million confirmed cases with a reported 2.4 million deaths. Both of these statistics are still climbing, causing more and more scholars to explore ways to help reduce the spread. A key to spread reduction is early detection of the virus. As of now, the most common way of testing for COVID-19 is using the reverse transcription-polymerase chain reaction (RT-PCR) test. The reason for this test being the most effective is because the virus has many symptoms similar to other, less-deadly viruses, such as the common cold. There are several issues with using RT-PCR. The time it takes for the test to come back is around 6-8 hours, making it a time-consuming test. In this processing time, the patients usually await the results in the Emergency Department, leaving a high risk of spread to patients and staff in the vicinity. Another issue is the expensive equipment that must be available in order to process the results. Along with these two issues, there presents the issue of false negative readings in the first stages of the virus. Finding a quick and less expensive method for diagnosis would be a tremendous help to the medical community. In order to help reduce the spread of the virus, a quick and easy diagnosis is important to procure. Machine learning can prove useful in accurately predicting COVID-19 test results. Using the data set provided by Dr. Langer and Dr. Favarato in their article, [4], the goal of this paper is to compare different supervised machine learning methods with the open-source WEKA environment [5] The different methods will be evaluated against baseline values, which are the F-measure percentages obtained in [4].

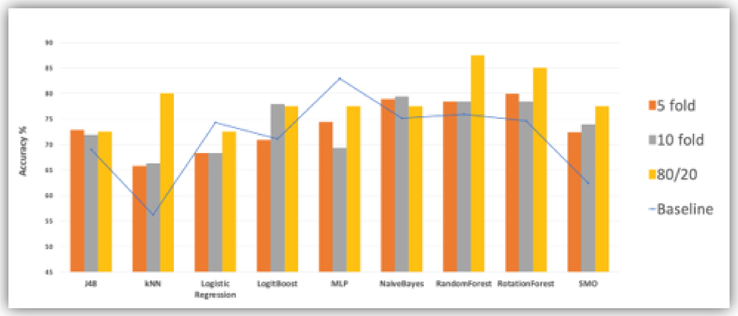
Research Question(s)

- How can machine learning be used to achieve better results using the same datasets from the Langer research?
- Does the TWIST method make a difference in helping to achieve better results? And if so, how much of a difference?
- Are the new results significant enough to make a difference in proving machine learning can help in the fight against COVID-19?

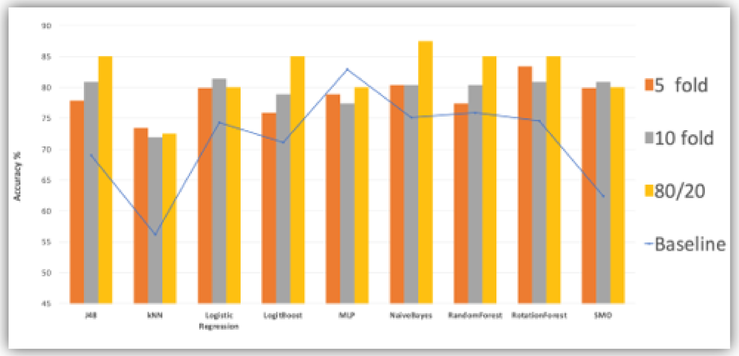
Materials and Methods

We decided to use a single machine learning language, WEKA, to come up with significantly better results achieved in the baseline testing. We used many different methods and techniques to achieve our top line results as described in the results section to the right.

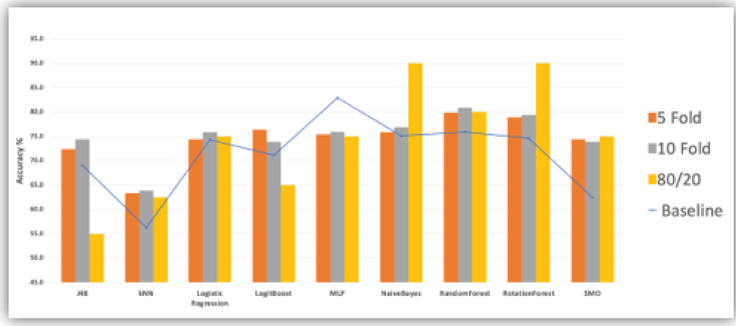
Results



Experiment #1 - Highest accuracy was recorded at 87.5 with the RotationForest algorithm using 80/20 Split for Cross Validation of the original 81-attribute dataset.



Experiment #2 - An accuracy of 85% was achieved with several models after using Feature Selection and 80/20 split for cross validation with the highest accuracy at 87.5% from NaïveBayes



Experiment #3 - The highest accuracy of the three experiments were achieved using the 41 TWIST attributes published in the original paper with a 90.0% accuracy using NaïveBayes and RotationForest with 80/20 Percent Split Cross Validation.

Conclusions

By using different test options and changing some settings we were able to achieve better results then were achieved in the original baseline using the same datasets. Best original baseline accuracy using WEKA was 82.9% on a 41-attribute dataset created using the TWIST method . As you can see from our research, we were able to achieve better results (87.5% accuracy) using the full 81-attribute dataset and an even higher result (90% accuracy) using the TWIST dataset. This leads us to believe that even though the TWIST method did lead to some improvement in the results, by further enhancing the methods and techniques used in WEKA we were able to see the greatest improvement over the original baseline results.
.

Acknowledgments

Special thanks to Dr. Seyedamin Pouriyeh for his invaluable knowledge and guidance on this project.

Contact Information

Project Team:
Bradley Durden (team leader) - bdurden4@students.kennesaw.edu
Mathew Shulman - mshulma1@students.kennesaw.edu
Andy Reynolds - freynol3@students.kennesaw.edu
Thomas Phillips - tphill71@students.kennesaw.edu
Indya Andrews - iandrew5@students.kennesaw.edu
Demontae Moore - dmoor195@students.kennesaw.edu

Advisor: Dr. Seyedamin Pouriyeh - spouriye@kennesaw.edu

References

[1] CDC.Aboutcovid-19.Sept.1,2020.[Online].Avail-able: https://www.cdc.gov/coronavirus/2019-ncov/cdcresponse/about-COVID-19.html

[2] ——.Symptomsofcoronavirus.Dec.22,2020.[On-line]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[3] W. H. O. Organization. Who coronavirus disease (covid-19) dashboard.Feb. 18, 2021. [Online]. Available: https://covid19.who.int/

[4] T. Langer, M. Favarato, R. Giudici, G. Bassi, R. Garberi, F. Villa, H. Gay,A. Zeduri, S. Bragagnolo, A. Molteniet al., "Development of machinelearning models to predict rt-pcr results for severe acute respiratorysyndrome coronavirus 2 (sars-cov-2) in patients with influenza-likesymptoms using only basic clinical data,"Scandinavian journal oftrauma, resuscitation and emergency medicine, vol. 28, no. 1, pp. 1–14, 2020.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H.Witten, "The weka data mining software: an update," ACM SIGKDDExploration Newsletter, vol. 11, November 2009 update,"ACM SIGKDDExploration Newsletter, vol. 11, November 2009.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Bradley Durden, Mathew Shulman, Andy Reynolds, Thomas Phillips, Indya Andrews, Demontae Moore
Advisor: Dr. Seyedamin Pouriyeh