

RESEARCH ARTICLE

A Predictive Model for Novel Corona Virus Detection with Support Vector Machine

Tsehay Admassu Assegie

Department of Computing Science, Faculty of Computing Technology, Aksum Institute of Technology, Aksum University, Aksum, Ethiopia

Received on: 15-01-2021; Revised on: 01-02-2021; Acceptance on: 16-03-2021

ABSTRACT

In this research, we have proposed machine-learning model for predicting coronavirus by employing support vector machine. The proposed model assists medical experts in the early detection of novel coronavirus disease (COVID-19) and helpful to make better decision during the coronavirus diagnosis. The objective of this study is to solve the problems caused by the wide-spreading coronavirus pandemic through machine learning approaches assisting the medical experts in detecting a novel coronavirus in the early stage. The detection of the coronavirus in early stage is vital in reducing the number of new cases of coronavirus through isolation of infected person from uninfected people. A support vector machine trained on the COVID-19 dataset collected from the online Kaggle data repository for coronavirus tested case dataset. Finally, we have analyzed the performance of the proposed model on coronavirus prediction with accuracy, precision recall, and receiver operating characteristic curve as performance metric in the evaluation of the model. The experimental test result reveals that the model has an accuracy score of 96.68% on detection of coronavirus.

Key words: Novel corona virus, Corona virus disease-19, Support vector machine, Machine learning, Disease prediction

INTRODUCTION

A novel coronavirus disease (COVID-19) is a family of virus that infects the human respiratory and hepatic organs which eventually causes death.^[1] The coronavirus is wide spreading all over the world causing death to human beings almost in all countries of the world. The pandemic is wide-spreading daily and the fear increased day by day due to the increased transmission rate.^[1,2] Combating the problem and the effect of COVID-19 on society requires much effort and research from multidiscipline such as healthcare experts, statistics, and computer science. Recently, computer science has become vital to the medical field due to the improved precision and faster speed in disease detection, treatment, and diagnosis. The challenge with COVID-19 is that transmission rate is higher compared to other diseases and there is no vaccine to the virus.^[2]

In addition to that, coronavirus is transmitted easily through handshake, kissing, touching the same surface, and any form of physical contact with infected person. This pandemic is causing thousands of deaths due to lack of proper medication, detection, and isolation in an early stages because infected people does not show symptoms for several days or even weeks. Hence, the detection of coronavirus is complicated because the number of medical experts is limited and the new cases introduced day by day are growing rapidly. The limited number of medical experts and equipment required for detection of coronavirus especially, in poor countries such as Ethiopia makes the identification of the virus more complicated. The objective of this research is to combat COVID-19 pandemic by developing predictive model that can assist the medical experts in detection of COVID-19 with machine learning algorithms. Furthermore, this study explores the answer to the following research questions: (1) How can we train a machine to detect COVID-19 in an early stage? (2) What is the relationship among novel coronavirus dataset

Address for correspondence:

Tsehay Admassu Assegie,
 E-mail: tsehayadmassu2006@gmail.com

features? (3) How can we combat COVID-19 with machine learning?

A massive increase in the number of novel coronavirus patients during the pandemic, has increased the workload on healthcare professionals. In addition to a very high workload on healthcare professionals, the risk for being infected during the diagnosis process has psychological impact on the healthcare professionals. The novel coronavirus pandemic have a negative impact on the society. Hence, reducing the risk with this pandemic requires an accurate and more precise automated predictive model that assists the health professionals' workload and stress due to the overwhelming cases of the pandemic.

RELATED WORKS

A limited number of works have been conducted on the range of applications of machine learning to medical challenges of the prediction of coronavirus. Some of the researches are discussed in this section. In Vaishya *et al.*,^[2] a predictive model for coronavirus patients' recovery is proposed with support vector machine, decision tree, and Naïve Bayes. The proposed model has an acceptable level of accuracy on predicting whether a coronavirus patient will survive or die. In Swapnarekha *et al.*,^[3] the role of intelligent computing to combat the coronavirus is reviewed. The authors reviewed machine learning application to automation of the coronavirus diagnosis using predictive model is crucial to combat the coronavirus problems as the identification of the

virus with automated intelligent systems reduce the time and the cost associated with diagnosis and identification of coronavirus.

Automation of the coronavirus diagnosis and identification with machine learning model plays significant role in combating the coronavirus pandemic.^[4,5] This is because automated diagnostic system are faster and cost-effective and have higher precision as compared to manual diagnosis. In Dansana *et al.*,^[4] the authors developed decision tree-based coronavirus prediction model. The experimental analysis on the performance of the proposed model shows an accuracy of 60% for the decision tree.

In Salehi *et al.*,^[6] the authors reviewed the performance of machine learning and deep learning model for automated coronavirus detection. The extensive review on machine learning and coronavirus shows that machine learning plays assistive role in combating coronavirus with acceptable level of performance [Figures 1 and 2].

RESEARCH METHOD

In this research, support vector machine-based coronavirus prediction model is developed using python programming language. The predictive model is trained on different observations of coronavirus dataset collected from online Kaggle data repository. The dataset consists of 2000 samples of which 913 are tested positive samples and 1087 samples are negative test samples. Each observation consists of a set of features, which are the symptoms of a coronavirus such as age,

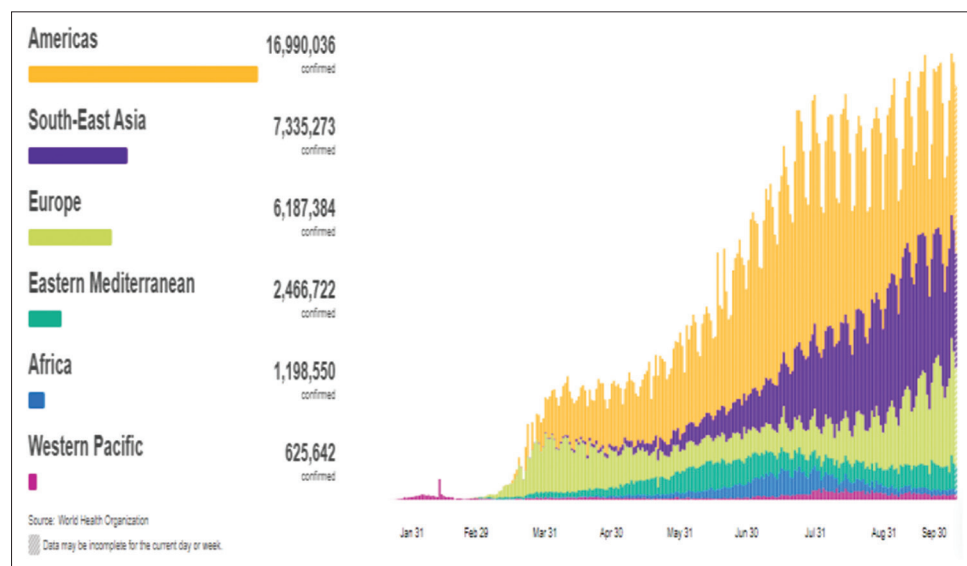


Figure 1: Corona virus cases October 4, 2020 (WHO)

headache, and shortness of breathing, sore throat, fever, and cough. The features taken as an input to the support vector machine and target feature, considered as the output of the model. In addition to automation of the coronavirus diagnosis with the use machine learning model, this research investigates the feature that has significant relationship to the target feature. The correlation between each feature in the coronavirus dataset is explored with Pearson correlation matrix as demonstrated in Figure 3.

As demonstrated in Figure 3, the coronavirus target feature has strong relationship with headache. This implies that the probability that a

person is tested positive for coronavirus is high if the person has headache. The features such as sore throat and fever have strong relationship with the target feature.^[7-9]

RESULT AND DISCUSSION

The experimental results are analyzed to validate the performance of the proposed coronavirus prediction model. Some of performance metrics used in evaluation include receiver operating characteristics (ROC) and accuracy. The performance of the proposed model experimentally

test_date	cough	fever	sore_throat	shortness_of_breath	head_ache	corona_result	age_60_and_above	gender
3/22/2020	1	0	0	0	0	0	0	1
3/22/2020	1	1	0	0	0	0	0	1
3/22/2020	0	1	0	0	0	0	1	1
3/22/2020	1	1	0	0	0	1	1	1
3/22/2020	1	0	0	0	0	0	1	0
3/22/2020	1	1	0	0	0	1	1	1
3/22/2020	0	0	0	0	0	0	0	1
3/22/2020	1	1	1	0	1	0	0	0
3/22/2020	1	0	1	0	0	0	0	0
3/22/2020	1	1	1	1	0	0	0	0
3/22/2020	1	0	0	0	0	0	0	1
3/22/2020	1	0	0	0	0	0	0	0
3/22/2020	1	0	0	0	0	0	0	0
3/22/2020	0	0	0	0	0	0	0	1
3/22/2020	0	0	0	0	0	0	0	0
3/22/2020	0	0	0	0	0	0	0	1
3/22/2020	1	1	0	1	0	0	0	1
3/22/2020	0	1	0	0	0	1	0	0
3/22/2020	0	1	0	0	0	1	0	0

Figure 2: The first twenty coronavirus dataset features used in training

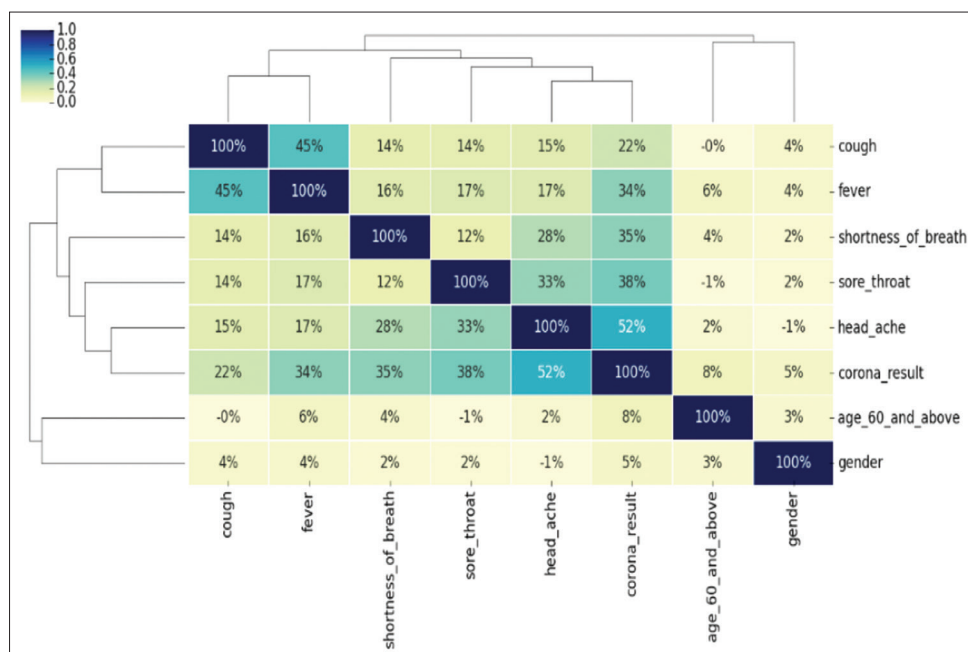


Figure 3: Coronavirus feature relationship

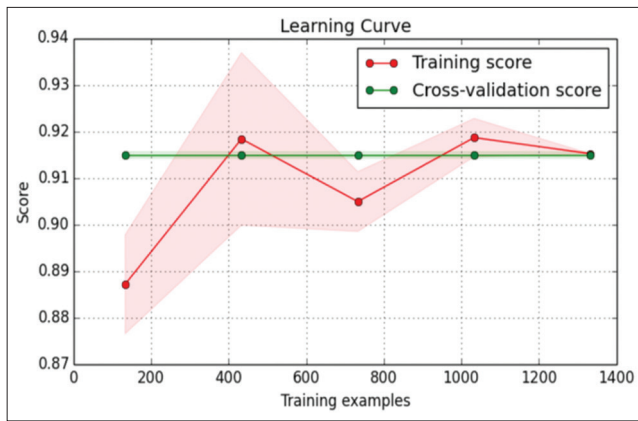


Figure 4: Learning curve of support vector machine

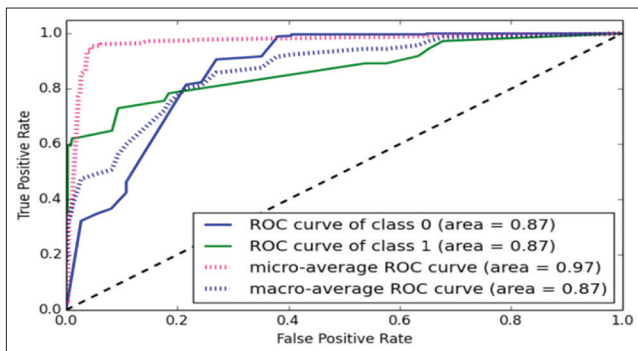


Figure 5: receiver operating characteristic curve of support vector machine

tested on the coronavirus dataset collected from online Kaggle data repository.

The learning curve of the model is demonstrated in Figure 4 for coronavirus dataset. As shown in Figure 4, the training score is low when smaller sample size is used and is maximum when roughly 1000 training samples are used. The cross-validation score remained constant.

As demonstrated in Figure 5, the area under the ROC curve is 0.87, which implies the overall accuracy of the model is acceptable. However, the model has better performance on predicting the class 0 or true negative class as compared to the class 1 or true positive class.

CONCLUSION

Coronavirus has caused many problems on the society in terms of life and economic activity. Hence, the application of machine learning to the diagnosis of coronavirus plays a significant role

in minimizing the effect of the coronavirus on the society. The automation of coronavirus diagnosis plays a great role in minimizing the time required for diagnosis by medical experts. In addition to that, automated diagnosis of coronavirus is vital to support the decisions made by the medical experts providing accurate and precise results as compared to human experts. The implementation of automated coronavirus prediction model presented in this study ensures quicker diagnosis with limited number of medical experts and better decision-making during the diagnosis of the coronavirus. The performance of the proposed model is evaluated on test set and result shows that 96.68% accuracy is achieved.

REFERENCES

1. Muhammad LJ, Islam M, Usman SS, Ayon SI. Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery, SN Computer Science; 2020.
2. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) Applications for COVID-19 pandemic, Diabetes and Metabolic: Clinical Research and Reviews; 2020.
3. Swapnarekha H, Beheraa HS, Nayak J, Naik B. Role of Intelligent Computing in COVID-19 prognosis: A State-of-the-Art Review. Netherlands: Elsevier; 2020.
4. Dansana D, Kumar R, Bhattacharjee A, Hemanth DJ, Gupta D, Khanna A, Castillo O. Early Diagnosis of COVID-19-Affected Patients Based on X-Ray and Computed Tomography Images using Deep Learning Algorithm, Soft Computing; 2020.
5. Kassania SH, Kassasib PH, Wesolowskic MJ, Schneidera KA, Deters R. Automatic Detection of Coronavirus Disease (COVID-19) in X-Ray and CT Images: A Machine Learning Based Approaches; 2020.
6. Salehi AW, Baglat P, Gupta G. Review on Machine and Deep Learning Models for the Detection and Prediction of Coronavirus. Netherlands: Elsevier; 2020.
7. Breast Cancer Prediction Model with Decision Tree and Adaptive Boosting. Vol. 10. IAES International Journal of Artificial Intelligence (IJ-AI); 2021.
8. Assegie TA, Nair PS. The performance of different machine learning models on diabetes prediction. Int J Sci Technol Res 2020;9:2491-4.
9. Assegie TA, Sushma SJ, Kumar SC. Weighted Decision Tree Model for Breast Cancer Detection. Vol. 62. Osaka: Technology Reports of Kansai University; 2020.