

A PROPOSED MODEL CAN CLASSIFY THE COVID-19 PANDEMIC BASED ON THE LABORATORY TEST RESULTS

S. Yasar and C. Colak

Abstract— As reported by the World Health Organization (WHO) in March 2020, COVID-19 is a worldwide epidemic. Since the rapid spread of the epidemic harms humans, the need for methods that enable early diagnosis and treatment has increased. Machine learning (ML) methods can play a vital role in identifying COVID-19 patients. In this study, the classification algorithms of ML methods (CART), Support Vector Machine (SVM-Radial), K Nearest Neighbors (K-NN) and Random Forest are used to determine the best model that diagnoses COVID-19 from the person's complete blood counts (positive/negative). According to the experimental results, the Random Forest algorithm gives the best predictions in the classification of COVID-19 (99.76% of accuracy). Besides, in the classification of Covid-19, it can be recommended to apply meta-learning algorithms as they can give high predictive results.

Keywords— COVID-19, Machine Learning, CART, SVM, K-NN, Random Forest

1. INTRODUCTION

THE COVID-19 outbreak, caused by the Severe Acute Respiratory Syndrome (SARS-CoV-2) virus, which first appeared in Wuhan, China's Hubei province on December 31, 2019, has rapidly spread to hundreds of countries. The epidemic process, which started in our country with the identification of the first positive case on March 11, 2020, continues increasingly. This disease agent has become the most critical health problem of the 21st century due to its high contagious feature, its unfavorable clinical prognosis, and its lethal effect in almost every age group, especially those aged 65 and above. Therefore, since the isolation of the new type of coronavirus, researches on COVID-19 disease and SARS-CoV-2 virus have been started in many countries, and it is possible to provide accurate information about the disease with these studies. However, as of the current situation, there are many issues related to the COVID-19 disease that have not yet been clarified [1].


Although developments in the field of health and technology are pleasing, some uncertainties encountered in medicine make it difficult for clinicians to make decisions. In recent years, artificial intelligence (AI) and machine learning (ML)-based studies have gained importance in terms of supporting the

decision of doctors to solve problems such as the diagnosis of diseases in the field of health.

Artificial intelligence is the general name of the technology to create and develop machines that can exhibit human mental activities and behaviors without the support of a living being [2]. Machine learning (ML), a sub-branch of artificial intelligence (AI), is a branch of computer science that makes decisions using past experiences when it is necessary to make decisions for the future. It allows a model to learn automatically from experience based on data, without having to model it just like statistical models. ML creates a new rule from the given samples [3]. It is also relevant to other disciplines such as machine learning, artificial neural networks, pattern recognition, information retrieval, artificial intelligence, data mining, and function approach. Compared to these areas, machine learning focuses more on how to model, optimize, and get organized for the best use of accessible training data. Classification algorithms used when data is tagged in machine learning are also frequently used in this field [4]. The purpose of classification algorithms is to predict which of the pre-labelled data groups similar data belongs to.

Many different algorithms have been developed for the classification process in the literature. Some of these are Decision Trees (CART algorithm), Support Vector Machines (SVM-Radial), K- nearest neighborhood (K-NN) and Random Forest (RF) algorithms. In the decision trees algorithm, the prediction can be obtained by building a decision tree with test points and branches. Each test point of a decision tree involves testing a specific input variable, and each branch represents the decision being made. A node that does not contain more branches is called a leaf node. The depth of the node is the minimum number of steps required to get from the node to the root [5]. In support vector machines algorithm, a technique called kernel trick is used to transform data. Kernel trick methods determine the optimal boundary among possible outcomes based on data transformation models. That is, kernel solution methods first perform complex data transformations and then determine how this data is to be separated based on the defined classes or results [6]. K-Nearest neighbor algorithm is one of the most known and used algorithms in machine learning algorithms. In this algorithm, a selected feature is classified using the distance between its closest feature. The K value found here is expressed as a number such as 3 or 5, for example [7]. The working principle of the Random Forest algorithm creates multiple decision trees and combines them to obtain a more accurate and stable forecast. Each tree is trained with different training sets and uses the results obtained. Each tree produces a classifier. These produced classifiers vote among

Seyma YASAR, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (seyma.yasar@inonu.edu.tr) 

Cemil COLAK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr) 

Manuscript received Sep 21, 2020; accepted Nov 16, 2020.
Digital Object Identifier:

themselves, and the algorithm determines the classifier with the most votes. This selected classifier is used to classify the data when new data is given [8].

This study aims to examine the data set consisting of laboratory test results obtained from patients with negative or positive SARS-Cov-2 test results with learning methods [such as decision trees (CART), support vector machines (SVM-Radial), K nearest neighborhood (K-NN) and random forest] to predict the COVID-9. It is foreseen that the best classification method can be used as an auxiliary system to diagnose the COVID-19 disease. Besides, a prediction model that can provide diagnostic support to physicians and other healthcare professionals is proposed to classify the COVID-19 pandemic.

2. MATERIAL AND METHODS

2.1. Dataset

The data set used in this study is the anonymized version of the SARS-CoV-2 RT-PCR and complete blood count (14 different) results of 5644 patients who applied to Israelita Albert Einstein Hospital in São Paulo, Brazil as suspicious. Complete blood count data are standardized to have mean zero and unit standard deviation before sharing. Some variables in the data set were excluded from the analysis because the missing data rate was over 90%. Explanations regarding the variables and their properties used in the creation of algorithms are given in Table 1.

TABLE I
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
Class	SARS-Cov-2 exam result (negative/positive)	Qualitative	Output
HCT	Hematocrit	Quantitative	Predictor
HGB	Hemoglobin	Quantitative	Predictor
PLT	Platelets	Quantitative	Predictor
MPV	Mean platelet volume	Quantitative	Predictor
RBC	Red blood Cells	Quantitative	Predictor
LYM	Lymphocytes	Quantitative	Predictor
MCHC	Mean corpuscular hemoglobin concentration	Quantitative	Predictor
LEU	Leukocytes	Quantitative	Predictor
BASO	Basophils	Quantitative	Predictor
MCH	Mean corpuscular hemoglobin	Quantitative	Predictor

The class variable included in the study consists of a total of 601 samples, 83 positive samples and 518 negative samples. Since it is thought that this class imbalance problem will negatively affect the training and test performances of the algorithms, the R programming language has been balanced using the "unbalanced" package.

2.2. Decision Trees

Decision trees technique are one of the simplest forms of the decision model, and they use sample data properties to create

their rules, which are in the form of a tree structure. Classification of data using the decision tree technique is a two-step process, which is learning and classification. In the learning stage, previously known training data is analyzed by the classification algorithm in order to create a model. The learned model is shown as classification rules or decision trees. In the classification step, test data are used to determine the accuracy of the classification rules or decision tree. If the accuracy is acceptable, the rules are used to classify new data. It should be determined in which order which fields in the training data will be used to create the tree. The most widely used measurement for this purpose is the Entropy measurement. The greater the measure of entropy, the more uncertain and unstable the results obtained using that field. Therefore, fields with the least entropy measure are used in the root of the decision tree [9]. There are many decision tree algorithms. The Classification and Regression Tree (CART) used in this study is a binary decision tree created by dividing a variable into two consecutive nodes (by repeating the process until the homogeneity criterion is reached), starting from the root node containing two learning steps [10]. The essential step in the creation of the CART decision tree is that branching to classify variables in the data set to determining according to which the criteria or which variable to do. At this stage, the variable with the lowest uncertainty is processed and used for testing at the root node [11].

2.3. Support Vector Machines (SVM)

A support vector machine creates an n-dimensional hyperplane that optimally divides the data into two categories. SVM models are closely related to artificial neural networks, SVM using a sigmoid kernel function; It has a two-layer feed-forward neural network. SVM uses kernel functions with high dimensional properties to ensure high performance when performing classification. In this study, a radial based kernel function is used as kernel function [12].

2.4. K Nearest Neighbors Algorithm (K-NN)

The nearest neighbors algorithm is a robust and versatile classifier often used as a reference point for more complex classifiers such as SVMs. In the K-NN algorithm, the samples in the training set are specified with n-dimensional numerical attributes. All training samples are kept in an n-dimensional sample space, with each sample representing a point in n-dimensional space. When an unknown sample is encountered, k nearest samples are determined from the training set, and the class label of the new sample is assigned according to the majority vote of the class labels of the k nearest neighbors [13].

2.5. Random Forest

Developed by Breiman, the Random Forest classifier creates a community with many decision trees it creates [14]. With the Bootstrap sampling method, different subsets are created from the data set and each decision tree trains the decision tree in the community. The best of the randomly determined variables for each node is considered, and the nodes are branched [15]. In the Random Forest algorithm, classification is started by determining two main parameters by the user. One of these parameters is the number of decision trees planned to be created, and the other parameter is the number of variables that perform the split. The Random Forest classifier uses classification and regression trees (CART) method to generate

trees. In the classification process, the result of the estimation of each tree is taken, and the majority decides what the classifier will be of votes method [14]. Random Forest classifiers are fast working classifiers as well as being resistant to overfitting problem. Two-thirds of the data set is divided into training data set and the remaining part as test data. Using a large number of decision trees as classifiers and generating a general estimate with the majority vote from the estimates obtained from each classifier decreases the bias while reducing the error rate in the estimates [15].

3. RESULTS

Performance criteria for making a COVID-19 diagnosis (negative/positive) using decision trees (CART), support vector machines (SVM), K- nearest neighborhood (K-NN) and random forest algorithms used in the study and 95% confidence intervals for these criteria are given in Table II.

TABLE II

PERFORMANCE METRICS AND A 95% CONFIDENCE INTERVAL FOR THE DATASET

Classification Algorithm	Accuracy (%95 CI)	Sensitivity (%95 CI)	Specificity (%95 CI)	MCC (%95 CI)	F1-Score (%95 CI)
Decision Trees (CART)	81.20 (77.4-84.9)	80.95 (77.17-84.73)	81.46 (77.7-85.2)	62.41 (57.7-67.1)	81.34 (77.6-85.1)
SVM (Radial)	92.77 (90.2-95.2)	96.19 (94.35-98.03)	89.27 (86.2-92.2)	85.72 (82.3-89.1)	93.09 (90.6-95.5)
K-NN	90.60 (87.7-93.4)	100.00 (100-100)	80.98 (77.2-84.7)	82.64 (79-86.3)	91.50 (88.8-94.2)
Random Forest	99.76 (99.2-99.9)	100.00 (100-100)	99.51 (98.8-99.9)	99.52 (98.9-99.9)	99.76 (99.3-99.9)

MCC: Matthews correlation coefficient

TABLE III

THE VARIABLE IMPORTANCE SCORE FOR COVID-19 ACCORDING TO THE RANDOM FOREST

Variables	Importance Score
LEU	157.64
PLT	125.95
EOS	87.99
MONO	75.56
RBC	46.09
LYM	44.66
MPV	43.73
HCT	40.17
BASO	39.30
HGB	37.29
MCV	33.10
MCHC	32.75
RDW	32.00
MCH	31.76

Accuracy performance values for CART, SVM, K-NN and Random Forest algorithms used in the classification of COVID-19 disease were found as 81.20, 92.77, 90.60 and 99.76, respectively. Therefore, the algorithm with the best accuracy according to the performance criteria obtained from the algorithms used in classifying COVID-19 disease is the Random Forest algorithm. The importance scores of the complete blood count values for the COVID-19 classification according to the random forest algorithm are given in Table III.

4. CONCLUSION

The studies and reports published by the World Health Organization (WHO) on the new coronavirus, which has spread to the world in a short time since its emergence in Wuhan, China, are followed with interest and concern by the whole world [16]. Therefore, considering the harmful effects of the pandemic on humans, it is crucial to detect COVID-19 positive cases at early stages and to make an immediate and correct intervention. The symptoms of COVID-19 are similar to those of the common cold and flu, making early diagnosis difficult for clinicians.

In this study, different classification algorithms (CART, SVM, K-NN, Random Forest) that classify COVID-19 by using the complete blood count results of COVID-19 positive patients were created. When the experimental results are examined, the accuracy values for each model in the COVID-19 (positive/negative) classification from the person's complete blood count are higher than 81%. Therefore, the classification performance of each model created is remarkable. Finally, when the COVID-19 classification performances of the algorithms used are examined, the random forest algorithm has the best performance.

On the other hand, medical imaging techniques are alternative to the methods used to diagnose COVID-19. Therefore, models that allow early diagnosis and diagnosis with different machine learning algorithms have been developed over images obtained from imaging techniques such as Computed Tomography (CT) and Chest radiography (CXR).

In a study have been used different deep learning models (ResNet18, ResNet34, InceptionV3, InceptionResNetV2, DenseNet161, and their Ensemble) to classify COVID-19 using Chest X-Ray images from infected/non-infected subjects. Among the performance metrics obtained in classifying the COVID-19 of the created models, the average F1-Score varies between 0.66 and 0.875. On the other hand, the average F1-Score value of the model created by ensembling all models was found to be 0.89 [17].

In another study, AlexNet, VGG-16, VGG-19, SqueezeNet, GoogleNet, MobileNet-V2, ResNet-18, ResNet-50, ResNet-101 and Classification models using Xception convolutional neural networks are used. Among all models, ResNet-101 and Xception have the best performance. It can be said that ResNet-101 and Xception convolutional networks are quite successful in distinguishing COVID-19. ($AUC_{ResNet-101}= 0.994$, $Sensitivity_{ResNet-101}=100\%$, $Specificity_{ResNet-101}=99.02\%$, $Accuracy_{ResNet-101}=99.51\%$, $AUC_{Xception} = 0.994$,

Sensitivity_{Xception}= 98%, 04, Specificity_{Xception}= 100%, Accuracy_{Xception}= 99.02%). Considering all performance metrics, it can be said that the model created using the ResNet-101 convolutional network is successful in COVID-19 classification [18].

In another study, a new DarkNet model was created using the YOLO deep learning architecture, which classifies automated COVID-19 as binary (positive / no finding) and multi-class (positive / no finding/pneumonia) using raw chest x-ray images. The accuracy value of the created model was obtained as 98.08% for binary classification and 87.02% for multi-class classification [19].

A recent study has focused on predicting whether a person is COVID-19 (positive/negative) in the early stage of the disease using ANN, RF and glmnet classifier on solely from 14 different blood counts. Models created in the study using ANN, RF and glmnet classifier can classify COVID-19 disease with accuracy rates of 80%, 86% and 84%, respectively [20].

In another study, using the same data set, an experimental study was conducted to predict patients with negative or positive COVID-19 class using SMOTE and ANN models. While the accuracy value was found to be 86% for the ANN model applied before balancing the data set, the accuracy value was obtained as 90% for the ANN model applied to the data balanced with SMOTE [21].

As a result, there have been many reported studies concerning medical imaging techniques and deep learning models to diagnose the COVID-19. However, the current study proposes a random forest model to classify the COVID-19 based on the complete blood count results. The Random Forest algorithm gives the best predictions in the classification of COVID-19 (99.76% of accuracy). Besides, in the classification of Covid-19, it can be recommended to apply meta-learning algorithms as they can give high predictive results.

REFERENCES

- [1] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, *et al.*, "Coronavirus disease 2019 (COVID-19): a perspective from China," *Radiology*, p. 200490, 2020.
- [2] R. Mitchell, J. Michalski, and T. Carbonell, *An artificial intelligence approach*: Springer, 2013.
- [3] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347-1358, 2019.
- [4] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," in *NDSS*, 2015, p. 4325.
- [5] D. Dietrich, B. Heller, and B. Yang, "Data science & big data analytics: Discovering," *Analyzing, Visualizing and Presenting Data*, 2015.
- [6] E. Gündoğan, A. K. Arslan, and J. Yağmur, "Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama," *Firat Tıp Dergisi*, vol. 22, 2017.
- [7] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunali, M. Şimşek, and A. Özçift, "KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi," 2016.
- [8] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *International workshop on machine learning and data mining in pattern recognition*, 2012, pp. 154-168.

- [9] C.-F. Chien and L.-F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems with applications*, vol. 34, pp. 280-290, 2008.
- [10] J. Han and M. Kamber, "Data Mining Concepts and Techniques, Morgan Kaufmann Publishers," *San Francisco, CA*, pp. 335-391, 2001.
- [11] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14-23, 2011.
- [12] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 32, pp. 440-448, 2002.
- [13] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.
- [14] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.
- [15] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational statistics & data analysis*, vol. 52, pp. 2249-2260, 2008.
- [16] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *Journal of autoimmunity*, p. 102433, 2020.
- [17] S. Chatterjee, F. Saad, C. Sarasaen, S. Ghosh, R. Khatun, P. Radeva, *et al.*, "Exploration of Interpretability Techniques for Deep COVID-19 Classification using Chest X-ray Images," *arXiv preprint arXiv:2006.02570*, 2020.
- [18] N. Khadem and A. Mohammadi, "Application of Deep Learning Technique to Manage COVID-19 in Routine Clinical Practice Using CT Images: Results of 10 Convolutional Neural Networks."
- [19] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, p. 103792, 2020.
- [20] A. Banerjee, S. Ray, B. Vorselaars, J. Kitson, M. Mamalakis, S. Weeks, *et al.*, "Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population," *International immunopharmacology*, vol. 86, p. 106705, 2020.
- [21] M. Yavaş, A. Güran, and M. Uysal, "Covid-19 Veri Kümesinin SMOTE Tabanlı Örneklemeye Yöntemi Uygulanarak Sınıflandırılması," *Avrupa Bilim ve Teknoloji Dergisi*, pp. 258-264.

BIOGRAPHIES

Şeyma YAŞAR obtained her BSc. Degree in mathematics from Gaziosmanpaşa University in 2009. She received an MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues PhD degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Cemil ÇOLAK obtained his BSc. Degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and a Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in general image processing, artificial intelligence, data mining, analysis.