Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **21** (2021) 025102 (300-309) AKU J. Sci. Eng. **21** (2021) 025102 (300-309)

DOI: 10.35414/akufemubid.788898

Araştırma Makalesi / Research Article

COVID-19 Diagnosis Prediction in Emergency Care Patients using Convolutional Neural Network

Kemal ADEM[1], Serhat KILIÇARSLAN[2*]

[1] *Aksaray University, Department of Management Information Systems, Aksaray, Turkey*

[2] *Gaziosmanpasa University, Department of Informatics, Tokat, Turkey*

**Corresponding author e-mail: serhatklc@gmail.com, ORCID ID : https://orcid.org/0000-0001-9483-4425*
kemaladem@gmail.com, ORCID ID : https://orcid.org/0000-0002-3752-7354

Geliş Tarihi: 01.09.2020 Kabul Tarihi: 12.04.2021

Abstract

The sudden increase in cases of Coronavirus disease (COVID-19) puts a high pressure on health care providers in many countries across the world. In the present case, an early and correct diagnosis of the disease, and starting the treatment is of vital importance. Most of the developing countries have insufficient RT-PCR tests, the most verified diagnostic test for COVID-19. This increases the number of infected patients and delays preventive measures. In this study, the risk of a positive COVID-19 diagnosis is estimated by applying Convolutional Neural Network (CNN) method, which is a deep learning model, to the dataset obtained from routine blood tests of all patients who admitted to the emergency service. The dataset used in the experiments consists of the data from patients admitted to the Israelita Albert Einstein Hospital in São Paulo, Brazil, between March 28th and April 3rd, 2020. In addition to the J48, ANN, Random Forest, and Random Committee data mining algorithms, the CNN deep learning algorithm were applied to the dataset. The 5 and 7 fold cross validation model was applied to the data set and the average of the two models was used as an evaluation criterion in order to ensure objectivity. The best prediction performance was obtained by the CNN method by 92.52% accuracy. Experimental results revealed that the proposed approach is in line with the results of the tests with general validity.

Keywords
COVID-19;
Data mining;
Deep learning;
CNN

Acil Bakım Hastalarında Evrişimli Sinir Ağını Kullanarak COVID-19 Hastalığının Tahmini

Öz

Koronavirüs hastalığı (COVID-19) vakalarındaki ani artış dünya genelinde birçok ülkenin sağlık hizmetleri üzerinde yüksek bir baskı oluşturmaktadır. Mevcut durumda hastalığın erken ve doğru tanısının koyulup tedaviye başlanması hayati önem taşımaktadır. COVID-19 için en doğrulanmış tanı testi olan RT-PCR gelişmekte olan ülkelerin çoğunda yetersizdir. Bu durum enfekte olan hasta sayısını arttırmakta ve önleyici tedbirleri geciktirmektedir. Bu çalışma ile acil servise gelen tüm hastalardan rutin olarak alınan kan testlerinden elde edilen veri kümesine derin öğrenme modellerinden Evrişimsel sinir ağı (CNN) yöntemi uygulanarak pozitif COVID-19 tanısı riski tahmin edilmektedir. Deneylerde kullanılan veri kümesi Brezilya, São Paulo'de bulunan Israelita Albert Einstein hastanesine başvuran hastalardan 28 Mart – 3 Nisan tarihleri arasındaki alınan verilerden oluşmaktadır. Veri kümesine J48, YSA, Random Forest ve Random Comittee veri madenciliği algoritmalarının yanında CNN derin öğrenme algoritması uygulanmıştır. Veri kümesine 5 ve 7 katlı çapraz geçerlilik modeli uygulanarak objektifliğin sağlanması açısından iki modelin ortalaması değerlendirme ölçütü olarak kullanılmıştır. En iyi tahmin performansı olan %92.52 doğruluk değeri CNN yöntemi ile elde edilmiştir. Deneysel sonuçlar, önerilen yaklaşımın genel geçerliliği bulunan testlerin sonuçları ile paralellik gösterdiğini ortaya koymaktadır.

Anahtar kelimeler
COVID-19;
Veri Madenciliği;
Derin Öğrenme;
CNN

1. Introduction

In December 2019, a new outbreak of Coronavirus, believed to cause pneumonia, was detected in Wuhan city of Hubei province of the People's Republic of China, and inability to take this outbreak under control caused its spread to China, then Europe, and then the whole world, leading to the pandemic (Aktoz *et al*. 2020). The Coronavirus Disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2) species, has become an unprecedented public health crisis (Zhou *et al*. 2020; Lu *et al*. 2020). This disease is mostly seen in people in the 30-80 age group, and it is more severe in people over 60 years of age. It is very rare in young children, especially under the age of 15. Approximately half of individuals with the COVID-19 infection has insignificant or overlooked symptoms, while the other half has severe symptoms. The main symptoms are high fever, fatigue, dry cough, myalgia and dyspnea (shortness of breath) (Lu *et al*. 2020; Wu and McGoogan 2020). Approximately half of the patients with severe symptoms of the disease have other comorbidities, such as hypertension, diabetes, and cardiovascular diseases (Wu and McGoogan 2020). The Coronavirus Resource Center at Johns Hopkins University School of Medicine reported that the number of people with COVID-19 infection worldwide exceeded 4,000,000 as of May 13th, 2020, of which 300,000 resulted in mortality. On March 16th, 2020, White House officials called on global artificial intelligence researchers to develop text, data mining and deep learning techniques to cooperate with research institutes and technology companies to help research on COVID-19. In collaboration with leading research groups, the Allen Research Institute has published an open-source, weekly updated COVID-19 open research dataset that includes academic articles on COVID-19 to speed up new research projects that urgently require real-time data (COVID-19 Dataset, 2020). Thanks to large-scale dataset obtained from the COVID-19 patients, analyses are being carried out using advanced data mining algorithms to better understand the viral propagation pattern, to further improve diagnostic speed and accuracy, and to identify people potentially more susceptible to infection (Ge *et al*. 2020; Metsky *et al*. 2020; Randhawa *et al*. 2020; Yan *et al*. 2020; Erkan and Thanh 2019; Yavaş *et al*. 2020).

Studies that applied data mining algorithms related to COVID-19 disease were investigated. In the first of these studies, a research aimed at estimating the intensity of COVID-19 infection in Iran was carried out. In that study, linear regression and long- short-term memory (LSTM) models were used to estimate the number of positive COVID-19 cases, and both models were evaluated using 10-fold cross-validity, and the Root Mean Squared Error (RMSE) value was used as the performance criteria. As a result, the RMSE values, obtained by applying linear regression and LSTM models to the available dataset, were calculated as 7.56 and 27.18, respectively. Thus, it was revealed that the trends of outbreaks can be predicted using data mining algorithms (Ayyoubzadeh *et al*. 2020). In a different study, the number of cases in India until March 30th, 2020 were analyzed for making predictions for the next 2 weeks. The SEIR (Susceptible, Exposed, Infectious, Recovered) and regression models were applied to the dataset collected from the data pool provided by Johns Hopkins University between January 30th and March 30th, 2020. The performance of the data mining algorithms was evaluated with Root Mean Squared Logarithmic Error (RMSLE), which was 1.52 for the SEIR model, and 1.75 for the regression model. In addition, the spread value of the disease, R_0, was calculated as 2.02. According to the estimations, it was predicted that the number of cases for the 2-week period could increase by 5000-6000 (Pandey *et al*. 2020). In another study, the Naive Bayes data mining algorithm was used to classify the positive or negative COVID-19 virus infections in South Jakarta, Indonesia. As a result of the application of this method, positive cases were classified by 55.48% accuracy (Santoso *et al*. 2020). In another study that used the same dataset we used in this study, neural networks, support vector machine, random forest and logistic regression data mining algorithms were applied to the dataset created with the results from emergency service examinations to predict the risk of positive COVID-19 diagnosis. The best prediction performance was obtained by the support vector machine algorithm, by 85% specificity and 68% sensitivity values. Furthermore, the three most important variables affecting positive diagnosis were found to be the lymphocyte count, leukocyte count and eosinophil count, respectively (de Moraes Batista *et al*., 2020). In the study conducted by Castro, the classification performance with a support vector machine was examined to detect COVID-19 disease on blood analysis results obtained from Israelita Albert Einstein hospital. As a result of experimental studies, they reported that it gave the best performance with an accuracy of 89.1% (Castro, 2021) Tanaydin *et al*. conducted a study on the detection of COVID-19 using blood analysis samples obtained from Israelita Albert Einstein hospital. In the experimental study, they first reported that they obtained a new data set by clearing the missing 95% of the data set. By applying Logistic regression, XGBoost and SVM algorithms to the new data set, they reported that they achieved the best success rate of 92.31% with SVM (Tanaydin *et al*.2020). Schwab *et al*. examined the performance of algorithms to detect COVID-19 disease with logistic regression, artificial neural network, random forest and XGbost algorithms on blood analysis samples obtained from Israelita Albert Einstein hospital. As a result of experimental studies, they reported that it gave 66% accuracy performance with XGbost (Schwab *et al*.2020 Yao *et al*. examined the classification performance with SVM on a dataset with 28 attributes with 76 patients obtained from blood analysis to detect COVID-19 disease. They reported that they achieved a 79.26% success rate as a result of experimental studies (Yao *et al*. 2020).

Nowadays, deep learning methods are actively used in many areas such as classification, image processing, and prediction (Schmidhuber, 2015). 1D-CNN, one of the deep learning architectures, consists of several layers. Some of these are convolution, pooling, normalization, fully-connected and activation layers.

Disease diagnosis is based on the experience of doctors. Since the experience and knowledge of each doctor is different, differences can be shown in deciding the disease. For this purpose, it is important to develop decision support systems in terms of helping doctors in the diagnosis and treatment of the disease (Shaker *et al*., 2019; Virgeniya *et al*., 2020; Sanderson *et al*., 2020; Adem *et al*., 2019; Adem, 2018). In this study, it was aimed to classify the COVID-19 disease with CNN, which is one of the deep learning methods, by using the data set consisting of blood analysis values.

In this study, a one-dimensional CNN method, one of the deep learning models, was applied in addition to the J48, ANN, Random Forest, and Random Committee data mining algorithms, to the COVID-19 dataset collected from the patients admitted to emergency service. Application of deep learning models aims to increase the probability of diagnosis of patients at risk of COVID-19. In this way, it is aimed to optimize the decisions about the COVID-19 test priorities especially in developing countries.

2. Materials and Methods

2.1. Material

In the study, the classic data mining and deep learning model were applied on a one-dimensional digitized dataset in order to detect COVID-19 disease. The dataset used in the experiments consists of the data from samples from the SARS-CoV-2 RT-PCR and additional laboratory tests of the patients admitted to the Israelita Albert Einstein Hospital in São Paulo, Brazil. The dataset consists of the data collected between March 28th and April 3rd. Samples were obtained from the website "https://www.kaggle.com/einsteindata4u/covid19" . The COVID-19 clinical dataset used was normalized to have an average of zero and a unit standard deviation (Tanaydin *et al*.2020).

The original dataset contains 111 attributes and 5644 data, with a large number of missing data by 88%. It is possible to complete the missing data with different methods. However, filling in incomplete data on medical data is not the right approach in terms of obtaining realistic results (Banerjee *et al*., 2020). In the data set used in the study, the missing and incorrect data causes the methods used to decrease the predictive power, create difficulties in determining the hyper parameters of the algorithms and make the analysis of the study in general. Thus, they create inaccurate predictions, causing us to obtain invalid results (Kang, 2013; Little *et al*., 2012; Yan *et al*., 2020; Batista *et al*., 2020). For all these reasons, we removed the missing data from the dataset to get better results for each row and column. After removing the missing data, the dataset included 422 data, 19 attributes, and 2 classes. Of the total 422 COVID-19 test results in the data set, 361 were negative and 61 were positive.

Nineteen features were used for training the algorithms, including age_quantile, regular_ward, semi_intensive, UTI, Hematocrit, Hemoglobin, Platelets, Red blood cells, Lymphocytes, Mean corpuscular hemoglobin concentration (MCHC), Leukocytes, Basophils, Mean corpuscular hemoglobin (MCH), Eosinophils, Mean corpuscular volume (MCV), Monocytes, Red blood cell distribution width (RDW), Neutrophils and C-Reactive Protein mg/dL. Table 1 shows the descriptive statistical results of the features assigned for all patients according to COVID-19 diagnosis.

Table 1. Descriptive statistical results of the dataset used

Parameters	Mean	Std	Min	Max
age_quantile	11,9710	5	0	19
regular_ward	0,1114	0,3150	0	1
semi_intensive	0,0616	0,2407	0	1
UTI	0,0355	0,1854	0	1
Hematocrit	0,1171	0,9606	-4,0665	2,6627
Hemoglobin	0,1082	0,9560	-3,8444	2,6719
Platelets	-0,0622	0,9536	-2,0751	9,5320
Red blood Cells	0,0896	0,9898	-3,6356	3,6457
Lymphocytes	-0,0041	0,9213	-1,8310	3,2182
Mean corpuscular hemoglobin concentration (MCHC)	-0,0017	0,9915	-4,5356	3,3311
Leukocytes	-0,1186	0,7752	-1,9285	2,9498
Basophils	0,1099	0,7792	-1,1401	3,4417
Mean corpuscular hemoglobin (MCH)	0,0264	0,9168	-5,5194	2,0600
Eosinophils	0,0725	1,0578	-0,8355	8,3509
Mean corpuscular volume (MCV)	0,0340	0,9322	-5,1016	3,1506
Monocytes	0,0868	0,9372	-2,1637	4,5334
Red blood cell distribution width (RDW)	-0,0480	0,9513	-1,5981	6,9822
Neutrophils	0,0356	1,0126	-3,3398	2,5359
Proteina C reativa mg/dL	-0,1127	0,7720	-0,5354	5,7337

2.2. Method

The study proposes one-dimensional CNN, a deep-learning algorithm, and classical data mining algorithm to diagnose COVID-19 disease positivity and negativity, using a digitized one-dimensional dataset.

2.2.1. Convolutional neural network (CNN)

The convolutional neural network (CNN) model has been proposed by LeCun, inspired by the human brain (LeCun *et al*. 1990). CNN is widely used in many areas, such as image processing, signal processing, and classification, in particular. CNN has increased its popularity by revealing meaningful and hidden information from within the dataset used (Ciresan *et al*. 2010). The CNN model consists of one or more convolution layers, an activation layer, a normalization layer, a fully connected layer, and an output activation function. In this study, the classification of the disease was carried out using a one-dimensional COVID-19 dataset. The CNN architecture used in the study is shown in Figure 1.

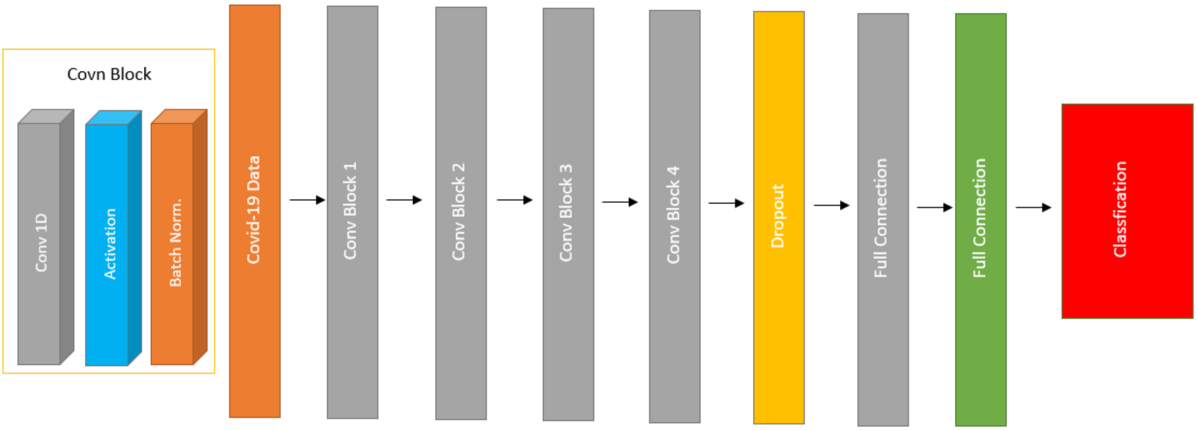

Figure 1. CNN architecture

In the proposed model shown in Figure 1, the convolution process, which is the first step of the CNN model, is applied to the dataset obtained after preprocessing the raw COVID-19 dataset. A 3x3 filter was applied by shifting on the proposed model in order to extract the low and high level features in the dataset determined in the convolution process. In this layer, weights were obtained by generating random numbers, and the convolution process was applied to the dataset with the resulting weights. This process continued with specific intervals until traversing the entire dataset. Finally, it resulted in the most significant feature map for the dataset. After the convolution process, the activation function was included in the feature map in the convolution layer. In the study, ReLU and LeakyReLU activation functions were used to enable better learning of the network. Following the activation, the non-normal distribution in the feature map was balanced by applying normalization (Deng and Yu 2014). After the normalization process, the dropout process was used to prevent memorization through a sparse network in the model. The dropout rate was chosen as 0.5 (Srivastava *et al*. 2014). After the dropout layer, the 256-neuron fully connected layer was used to connect all layers in order to perform a better classification, based on the similarities and differences of the features. This model performs the classification process that is the diagnosis of disease through the artificial neural network in this layer. Classification is performed using softmax, the output activation function (LeCun *et al*. 1990; LeCun *et al*. 1998; LeCun *et al*. 2015; Simard *et al*. 2003). While there are parameters to be defined in the conventional data mining algorithms, the CNN method provides a better assessment since it discovers the important parameters by itself. This feature emphasizes the superiority of deep learning algorithms over classical methods.

In this study, the CNN model was used to classify the data for the COVID-19 disease. In the proposed method, first a preprocessing was applied to the dataset. As for the preprocessing, missing data and features were removed from the raw dataset, since they were excessive and had a negative impact on the success rate. In the event of a large difference between the values of the parameters used as inputs, a normalization preprocessing is needed, which will normalize the input values and update them in a specific range. According to the experimental results, the most successful results were obtained by the proposed method using the min-max normalization technique, which normalizes the input values in the range of 0 to 1, in the data preprocessing step. The classification process was carried out by applying the proposed one-dimensional CNN model on the resulting dataset.

3. Discussion and Conclusions

Experimental studies were carried out using the dataset used in the study, classical data mining algorithms, and the deep learning model. For the j48 method, the batchSize value was set at 100 and the confidenceFactor value at 0.25. For the RF method, the batchSize value is set to 100 and the bagSizePercent value to 100. For the Random

Committee method, the numDecimalPlaces value is set to 2. J48 and Random Forest methods were preferred in the study because of the high performance and easy application of decision tree algorithms (Ali *et al*., 2012). The CNN structure was used as the deep learning model. The experiments of the study were carried out in the Weka and Matlab R2020a platforms. In the study, the application was performed on a system with the Intel Core i5 7200U 2.5 GHz processor, and 12GB DDR3 memory.

Table 2 shows the hyper-parameter ranges. The parameters for the 17-layer architecture used in the CNN model are presented in Table 3.

Table 2. Hyper-parameter range of architectures used

Hyperparameter	Range
Momentum	[0.7 - 1]
InitialLearnRate	[0.01 - 0.03]
LearnRateDropFactor	[0.1, 0.2, 0.3]
LearnRateDropPeriod	[20, 30, 40, 50]
Max-Epoch	[50,100,150,200]
Optimizer	[ADAM, SGD, RMSprop, Nadam]
Loss Function	[Categorical_crossentropy, Sparse_categorical_crossentropy]
Activation Function	[ReLU, LeakyReLU, ELU]
Output Activation	[Sigmoid, Softmax]

Table 3. CNN architecture and parameters used in the study

CNN Layers	Parameters		
Input Data	422x1		
Conv1	24@3x3 filter	Stride=2	Padding=same
Activation	ReLU-leakyReLU		
BatchNormalization			
Conv2	36@3x3 filter	Stride=2	Padding= same
Activation	ReLU-leakyReLU		
BatchNormalization			
Conv3	48@3x3 filter	Stride=2	Padding= same
Activation	ReLU-leakyReLU		
BatchNormalization			
Conv4	64@3x3 filter	Stride=2	Padding= same
Activation	ReLU-leakyReLU		
BatchNormalization			
Dropout	0.5		
Fully connected	256		
Fully connected	256		
Output Activation	Softmax		

Table 4 shows the hyper-parameters of the CNN model used in the study. 10 different experiments were carried out by determining the spacing values for the hyperparameters used in the CNN architecture. The most successful 2 models and their parameters are shown in Table 4.

Table 4. Hyperparameters of the CNN models

	Model 1	Model 2
Max-Eposch	200	100
Activation	ReLU	LeakyReLU
Momentum	0.9	0.78
InitialLearnRate	0.01	0.01
LearnRateSchedule	Piecewise	Piecewise
LearnRateDropPeriod	40	30
LearnRateDropFactor	0.1	
Loss Function	Categorical cross entropy	
solverName	SGDM	
activation	Softmax	

As shown in Table 3, the classification process was performed after applying 4 different convolutions, activation and normalization processes on the COVID-19 dataset. A 3×3 filter was used in the convolution layers, and a small stride value was selected to minimize data loss between layers. The learning rate was set to 0.01 in both models and stochastic gradient descent with momentum (SGDM) was used for the optimizer. In addition, a better learning model was aimed by gradually reducing the learning rate using piecewise. The accuracy, precision, sensitivity, F-score, and Kappa score values obtained by applying the proposed CNN deep neural network model and J48, ANN, Random Forest, and Random Committee classical data mining algorithms to the COVID-19 dataset used in the study are presented in Tables 5 and 6.

Table 5. Experimental results of the classical data mining models

Model	Train-Test	Accuracy (%)	F-Score (%)	Sensitivity (%)	Precision (%)	Kappa
J48	5 fold	86.23	81.95	83.42	82.12	0.8548
	7 fold	86.51	86.12	86.58	85.74	0.8586
	Mean	86.37	84.035	85	83.93	0.8567
ANN	5 fold	86.14	84.52	85.16	84.18	0.8529
	7 fold	88.22	87.19	88.35	87.56	0.8692
	Mean	**87.18**	**85.85**	**86.75**	**85.87**	**0.8610**
Random Forest	5 fold	83.87	80.12	82.66	80.79	0.8263
	7 fold	86.45	85.46	86.35	85.24	0.8514
	Mean	85.16	82.79	84.505	83.015	0.8388
Random Comittee	5 fold	82.54	79.52	82.50	81.63	0.8165
	7 fold	86.16	84.22	87.75	85.85	0.8532
	Mean	84.35	81.87	85.12	83.74	0.8348

As shown in Table 5, experimental studies on the COVID-19 disease dataset showed that the best success rate average was achieved by the artificial neural network (ANN) algorithm that number of neurons 20, learning rate of 0.3 and the momentum value of 0.2. Experimental results showed a success by 87.18% accuracy, 85.85% F-score, 86.75% sensitivity, 85.87% precision, and 0.861 Kappa score.

In Table 6, a digitized COVID-19 was studied with the CNN model, one of the deep learning methods. Following the application of the algorithms, the results of the classification of the digitized COVID-19 dataset are seen.

Table 6. Experimental Results of the CNN Deep Learning Method

Model	Train-Test	Accuracy (%)	F-Score (%)	Sensitivity (%)	Precision (%)	Kappa
Proposed CNN Model 1	5 fold	91.85	91.84	92.46	91.98	0.9084
	7 fold	92.45	91.79	92.78	92.08	0.9163
	Mean	92.15	91.82	92.62	92.03	0.9124
Proposed CNN Model 2	5 fold	92.25	92.05	93.67	92.12	0.9185
	7 fold	93.02	92.84	93.89	92.24	0.9291
	Mean	**92.64**	**92.45**	**93.78**	**92.18**	**0.9238**

Table 6 shows the success rates of the CNN model applied to the digitalized COVID-19 dataset in the study. The model used in the experimental study consists of 19 layers, including the input layer. Experiments were carried out by applying the 5 fold and 7 fold cross validation method for the two models established. In the study, it was found that the best success rate average was obtained by the LeakyReLU activation function and the momentum value of 0.78 for the proposed CNN_model_2. In addition, the experimental results show a 92.64% accuracy, 92.45% F-score, 93.78% sensitivity, 92.18% precision, and 0.9238 Kappa score. Detection studies performed on blood test data of different COVID-19 diseases in the literature using the same data set, data size, number of features, method and accuracy results are given in Table 7.

Table 7. Comparison of the methods for the classification of COVID-19 disease

Authors	Year	Data Size	Number of Classes	Number of Attributes	Method	Accuracy %
Batista et al.	2020	236	2	15	SVM	85
Santoso et al.	2020	249	2	11	Naive Bayes	55.48
Castro et al.	2020	204	2	-	SVM	89.1
Tanaydin et al.	2020	581	2	17	Logistic Regression, XGBoost and SVM	92.32
Schwab et al.	2020	5644	2	111	Logistic Regression, ANN, Random Forest XGBoost and SVM	66
Yao et al.	2020	76	2	28	SVM	79.26
Proposed CNN Model	2021	422	2	19	CNN	92.64

As can be seen in Table 7, in different studies using the data set used in the study, both the data and the attribute dimensions have been reduced and the methods used and the success results obtained are presented. When the results were examined, it was seen that the CNN model we suggested achieved higher success than other studies in the literature.

Results from all experimental studies show that CNN, a deep learning method, produces better results than classical data mining methods. The main reason for achieving the best success rate with CNN is that it increases the classification success by extracting the low and high level features in the dataset determined in the convolution process.

In the study, we can list the following points on the success of the CNN model.

- Among the data mining studies, the proposed CNN model was found to have better performance in identifying COVID-19 cases.
- The proposed CNN model is successful in detecting COVID-19 symptoms.
- It can help specialists for the diagnosis, follow-up, treatment and isolation of patients.
- Patients with COVID-19 positivity can be referred to more comprehensive centers for starting treatment as soon as possible.
- In addition, unnecessary chest X-rays and occupation of health care centers can be avoided for the patients diagnosed as negative by the model.

4. Conclusion

COVID-19 disease, which has emerged in Wuhan, China, is seen as a disease that threatens the world in a short time. In this respect, the treatment should be started quickly by making accurate and timely diagnosis. In our deep learning-based model in this study, COVID-19 cases can be checked and early stage assessments can be made by providing preliminary diagnosis about the pandemic disease, using basic blood tests from patients admitted to the emergency service. In this regard, a decision support system was developed to assess current blood test results using the CNN deep learning algorithm, for an accurate diagnosis and for helping physicians with their decisions. Experimental studies showed that the proposed CNN model has an average success rate of 92.64% on the COVID-19 disease dataset. As a result, the similarity between the COVID-19 test results and the values obtained by the application of data mining and deep learning algorithms to routinely collected blood data has paved the way for a promising new field. In this way, the developed data mining models can help policymakers and health care managers in planning the health resources and taking the outbreak under control. One limitation of this study is the use of blood data from a limited number of people. We aim to make our model more robust and accurate by using more blood test data from local hospitals. As a target study, we plan to conduct a time series analysis of COVID-19 data using the LSTM deep learning model. We also aim to make our model work mobile, in a way that health care professionals will use it for the pre-diagnosis of COVID-19.

5. References

Adem, K., 2018. Exudate detection for diabetic retinopathy with circular Hough transformation and convolutional neural networks. *Expert Systems with Applications*, **114**, 289-295.

Adem, K., Hekim, M., and Demir, S., 2019. Detection of hemorrhage in retinal images using linear classifiers and iterative thresholding approaches based on firefly and particle swarm optimization algorithms. *Turkish Journal of Electrical Engineering & Computer Sciences*, **27(1)**, 499-515.

Aktoz M., Altay H., Aslanger E., Atalar E., Aytekin V., Baykan A. O., Barçın C., Barış N., Boyacı A. A., 2020. COVID-19 pandemisi ve kardiyovasküler hastalıklar konusunda bilinmesi gerekenler. *Turk Kardiyol Dern Ars*, **48**, 1-87.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I., 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, **9(5)**, 272.

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., and Kalhori, S. R. N., 2020. Predicting COVID-19 ıncidence through analysis of google trends data in ıran: data mining and deep learning pilot study. *JMIR Public Health and Surveillance*, **6**, e18828.

Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., ... and Mackenzie, L. S., 2020. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. *International immunopharmacology*, **86**, 106705.

Castro, J. D. S., 2021. Discrimination of SARS-Cov 2 and arboviruses (DENV, ZIKV and CHIKV) clinical features using machine learning techniques: a fast and inexpensive clinical screening for countries simultaneously affected by both diseases. *medRxiv*.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J., 2010. Deep big simple neural nets for handwritten digit recognition. *Neural Computation*, **22**, 3207–3220.

de Moraes Batista, A. F., Miraglia, J. L., Donato, T. H. R., & Chiavegatto Filho, A. D. P., 2020. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*, 1-8.

Deng, L. and Yu, D., 2014. Deep learning: methods and applications. *Foundations and trends in signal processing*, **7**,197-387.

Erkan, U. and Thanh, D. N., 2019. Autism spectrum disorder detection with machine learning methods. *Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews*, **15**, 297-308.

Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., ... & Cheng, L., 2020. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. *bioRxiv*, 1-62.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, **64(5)**, 402.

LeCun, Y., Bengio, Y., & Hinton, G., 2015. Deep learning. *nature*, **521**, 436-444.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D., 1990. Handwritten digit recognition with a back-propagation network. *In Advances in neural information processing systems*, 396-404.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278-2324.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Stern, H., 2012. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, **367(14)**, 1355-1360.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., ... & Bi, Y., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus:

implications for virus origins and receptor binding. *The Lancet*, **395**, 565-574.

Metsky, H. C., Freije, C. A., Kosoko-Thoroddsen, T. S. F., Sabeti, P. C., & Myhrvold, C., 2020. CRISPR-based COVID-19 surveillance using a genomically-comprehensive machine learning approach. *bioRxiv*, 1-11.

Pandey, G., Chaudhary, P., Gupta, R., & Pal, S., 2020. SEIR and Regression Model based COVID-19 outbreak predictions in India. *arXiv preprint*, 1-10.

Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A., & Kari, L., 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos one*, **15(4)**, e0232391.

Sanderson, M., Bulloch, A. G., Wang, J., Williamson, T., & Patten, S. B., 2020. Predicting death by suicide using administrative health care system data: Can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance?. *Journal of affective disorders*, **264**, 107-114.

Santoso, P. H., Fauziah, F., & Nurhayati, N., 2020. Application of data mining classification for covid-19 ınfected status using algortima naïve method. *Jurnal Mantik*, **4**, 267-275.

Schwab, P., Schütte, A. D., Dietz, B., & Bauer, S. (2020). Clinical predictive models for COVID-19: systematic study. *Journal of medical Internet research*, **22(10)**, e21439.

Shaker, A. M., Tantawi, M., Shedeed, H. A., & Tolba, M. F., 2019. Heartbeat classification using 1D convolutional neural networks. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 502-511).

Simard PY, Steinkraus D, Platt JC., 2003. Best practices for convolutional neural networks applied to visual document analysis. *In Proceedings of the Seventh International Conference on Document Analysis and Recognition*, **2**: 958–962.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**, 1929-1958.

Tanaydin, A., Liang, J., & Engels, D. W., 2020. SARS-CoV-2 pandemic analytical overview with machine learning Predictability. *SMU Data Science Review*, **3(2)**, 17.

Virgeniya, S. C., & Ramaraj, E., 2020. Predictive Modeling Algorithms-based Classification of Arrhythmia. *In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1272-1276).

Wu, Z. and McGoogan, J. M., 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *Jama*, **323**, 1239-1242.

Yan, L., Zhang, H. T., Xiao, Y., Wang, M., Sun, C., Liang, J., ... & Tang, X., 2020. Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan. *medRxiv*, 1-14.

Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., ... & Wang, G., 2020. Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in cell and developmental biology*, **8**, 683.

Yavaş, M., Güran, A., & Uysal, M. Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, **(Special Issue)**, 258-264.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Chen, H. D., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, **579**, 270-273.