Prediction of COVID-19 Using Some Machine Learning Models and Its Comparison with a Deep Learning Model

Anam Naz Sodhar[1,*], Muhammad Awais Rajput[2], Saifullah Memon[3], Fizza Abbas Alvi[4], Irum Naz Sodhar[5], Abdul Hafeez Buller[6],

[1] Postgraduate Student, QUEST, Nawabshah, Pakistan
[2] Department of Artificial Intelligence, QUEST, Nawabshah, Pakistan
[3] State Key Laboratory of Networking and Switching Technology, BUPT, Beijing, China
[4] Department of Computer Systems Engineering, QUEST, Nawabshah, Pakistan
[5] Department of Information Technology, SBBU, Nawabshah, Pakistan
[5] Project Management Unit, QUEST, Nawabshah, Pakistan
*Corresponding author: anumakber10@gmail.com

Abstract

Coronavirus (COVID-19) started from Wuhan, China in December 2019. Since then, this virus has affected millions of people around the world and has caused deaths in millions. As of right now, there is no cure or permanent treatment for this disease. It is well known that machine learning plays an important role in the health care system. In this research, we are going to use some machine learning models such as Decision Tree (DT), logistic regression (LR), and Random Forest (RF) for the forecasting of corona virus. These models are implemented using different machine learning libraries available in Python. This work not only serves the purpose of COVID-19 predictions using machine learning, but also attempts to find out suitable model with the best features to save time and resources. Furthermore, we also compare some of the features of different machine learning models with a deep learning model (CNN). Since healthcare environment, computational resources have to be optimized, prediction models which use less computational resources are always preferred. We believe that the outcomes of this study can help understand the performance of various predictions models in the prediction of COVID-19.

Keywords—COVID-19, Machine Learning, Deep Learning, Random Forest, Decision Tree

1 Introduction

The first case of COVID-19 emerged in December 2019. This infection originated from SARS-CoV-2, a member virus of a large family known as coronavirus. This virus may also have originated in an animal and mutated so it can motive sickness in human beings. There is also a history of these viruses originating in different animals such as bats, pigs, birds, and other animals which mutated and became dangerous for the beings [1], [17]. Further research and study are ongoing to find the real cause of the pandemic.

The outspread of COVID-19 is incredible since it transmits from a person to other person. Coronavirus is getting spread through virus particles and a droplet released in the air when an infected person does an activity such as breathing, talking, coughs, sneezes, or singing. In a few seconds, large droplets may fall to the ground but small infectious particles accumulate in indoor places when there is a poor source of ventilation and many people are gathered [2], [18].

Most people infected with the coronavirus can suffer from mild to moderate respiratory illness and recover without any special treatment. While some end up severally ill and require clinical care. People with old age who are already going through conditions like diabetes, cardiovascular disease, chronic respiratory disease, or cancer are more to develop serious illnesses. The consequences include serious illness and even death. The best way to protect yourself and others from this virus is to be well informed about the disease and how it spreads [19].

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2001.12

As of right now, there is no cure or permanent treatment to get rid of this disease. For mild cases, doctors may recommend fever reducers or over-the-counter medications. While serious cases may require hospitalization where patients may be given steroids, mechanical breathing support with oxygen, and other COVID-19 treatments under development [1], [2]. Infusions of organism antibodies given to bound patients early within the infection might cut back the symptoms, severity, and period of sickness [17], [19].

Although there has been lately approved vaccination from multiple sources which can effectively work out to prevent an infection with SARS-CoV-2, but there is a lot to understand about their effects. Moreover, continued hygienic practice also helps to slow down the spread of COVID-19 [1]. Being vaccinated can only reduce the chances of getting infected but if you didn't keep the safety measures then you are at risk again.

The purpose of this study is to forecast or predict COVID-19 by applying some machine learning and a deep learning model. Further, this paper evaluates various measures for performance including the prediction accuracy, F1-score among others. Finally, the paper reports better models among the compared for the COVID-19 dataset.

2 Related Work

The investigation of COVID by using different techniques of Artificial Intelligence for is the need of the day, so the that reason this research study is based on the prediction of COVID-19 by using Machine Learning Models and Its Comparison with a Deep Learning Model. Many researchers worked different symptoms, texture and etc., but in [3] author discussed how a machine learning-based model is designed for the prediction of coronavirus based on a textual dataset. Predictions models were designed with the combination of several features to overcome the risk of disease. The goal of this was to help medical staff where resources are limited.

Authors in [4] discussed how COVID-19 was found in China and became a pandemic, and there is no curable medicine till now. This paper works out a deep learning model using CT images to predict virus presence. In [5] a review has been done on coronavirus infection and findings in this paper shows that machine learning has an important role in the prediction, investigation, and discrimination so machine learning can be utilized in the medical department, supervised learning shows better results as compared to unsupervised learning by having testing accuracy of 92.9%.

In another work [6] authors discussed how coronavirus is becoming an endemic, nearly 651,247 have lost their lives after getting infected from this disease, right now no cure or permanent treatment for this disease. It has become a burden on the healthcare system around the world where resources are limited especially in developing countries. In this work supervised machine learning models were developed using epidemiology datasets regarding cases in Mexico (positive/negative). They follow a typical split of train-test as 80-20%. In their performance evaluation, decision trees comes out to be the most accurate having 94.99% accuracy in comparison to SVM and Naïve Bayes Models [7].

In this research, we are going to use some Machine mastering models for the forecasting of coronavirus. In particular, we compare inclusive of Logistic Regression, Random Forest (RF), and Decision Tree (DT). These models are going to be applied with the aid of using python and one of a kind gadget getting to know libraries. The motive of this paintings is not just to do the predictions of COVID-19 however to find out the first-rate model with the first-class functions to keep time and assets.

3 Methodology

3.1 Collection of Dataset

The most important step in the process of data collection. This study utilizes COVID-19 dataset which has been obtained from Kaggle. The given twenty attributes in the dataset show the following dataset involves the information about different symptoms and any kind of another disease they are suffering from their history of other activities like abroad travel, contact with an infected person, attending large gatherings, and so on. This dataset is described in the numerical way class 0 and 1 (1 for positive or yes whereas 0 for no or negative).

3.2 Data Processing

After data collection the given data is cleaned first to remove the missing and duplicated values in case there are to bring the dataset in the same level of granularity, followed by data cleaning and afterwards splitting in to sets of training and testing [8].

3.3 Training

Next step involves applying training data to let the model learn the patterns inside the data. Following that the performance metrics can be evaluated to see how accurate the predictions are [8].

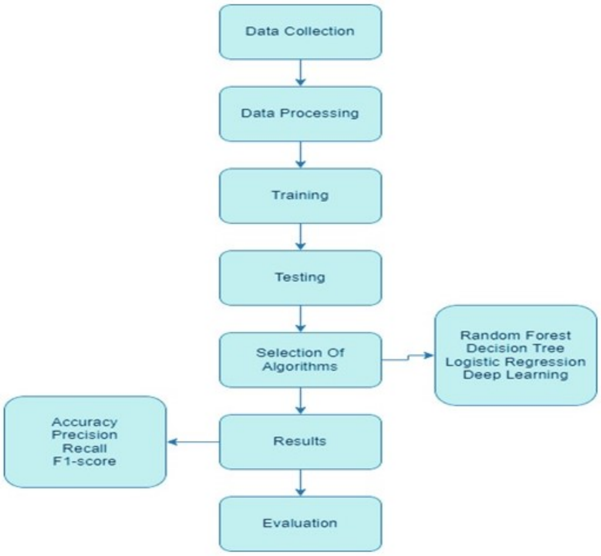

Fig. 1: Flow chart of the methodology

3.4 Testing

After the data has learned enough the model is evaluated on new data i.e., testing set. This makes sure that the model can predict on the data which it has never seen before. The evaluation on the testing data reveals its accuracy [8].

3.5 Selection of Algorithms

Classification is a supervised studying process that is used for predicting the consequence from present data. This work proposes a way for the analysis of coronavirus the usage of classification algorithms [9]. After the testing and training step, the next step is the selection of the algorithm. Some of the most commonly used models are Logistic Regression, Random Forests, and Deep Learning, among others.

3.5.1 Logistic Regression

A supervised algorithm which is used to predict the chance of a variable which is dichotomous, and potentially there would be solely two possible outcomes: Failure or success represented as 0 or 1 [9].

3.5.2 Decision tree

A supervised knowledge method that may be used for any type and regression problem; however, one usually needs to clear up type problems. It is a tree classifier, in which inner nodes suggest factors, branches represent desires policies, and the leaf node are the end-result. In a choice tree, there are nodes, which might be choice nodes and leaf nodes [10].

Decision nodes are employed mainly to represent features based on branches, whilst leaf nodes are the output of these choices and now not consist of comparable branches. Choices or assessments are made primarily based totally on components of the desired dataset. It is a graphical instance of having all viable alternatives for a problem/choice specifically below sure conditions.

It is referred to as a tree of desire because, just like a tree, it begins off evolved with the foundation node, which then grows into branches and constructs a tree. A popular technique is known as Tree Classification and Regression Algorithm i.e., CART algorithm. Primarily, a tree solves a question based on the reply (True or False) and in addition, cut up the tree into sub-trees.

3.5.3 Deep Learning

AI-based algorithms learn historical statistics to predict outcomes for unseen data. A commonly used subset of AI are and deep learning (DL) algorithms which primarily rely on complex learning models. DL algorithms have been characterized as computationally restrained as they require higher computing power and complexity. Yet, trends in massive records have made it possible to realize larger and more complicated networks to learn, examine, and react to complicated conditions quicker than humans. General areas where DL enjoy large share of applications include photo classification [11], speech recognition [12], and bioinformatics [13] etc.

Convolutional Neural Networks (CNNs) have recently claimed huge success in image classification tasks. Structurally its main element are convolutional layers followed by pooling layers, and a classification layer. Convolution mainly perform characteristic extraction. The pooling layer attempts to decrease the dimension of the inputs [14].

3.5.4 Random Forest

Another popular supervised algorithm for learning is Random forest (RF). The main elements in this algorithm are trees joining to form a forest. Each tree is responsible for a classification expectation and the category with the most votes are assembled and turned into the model's prediction [9]. The depth and variety of trees however, play an important role in the overall accuracy of the classifier. It is often termed equally efficient as regression task, however it is better suited at classification tasks, and more importantly can account for missing values.

3.6 Performance Measure

For performance measure accuracy, F1-score, recall, and precision are calculated.

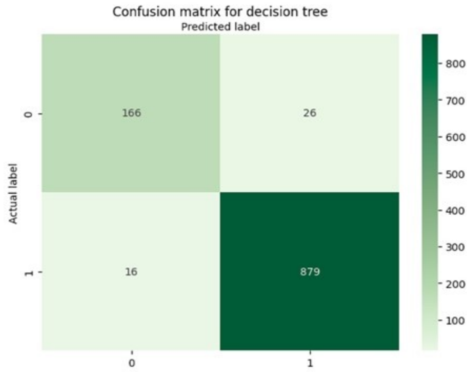

Fig. 2: Decision tree

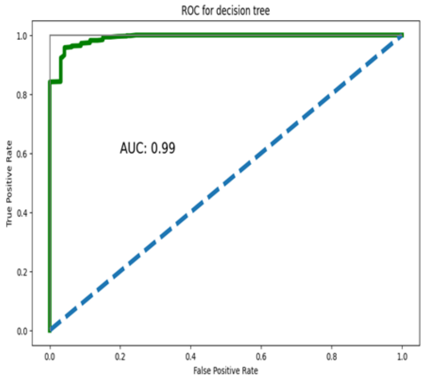

Fig. 3: ROC for Decision tree

Confusion Matrix: Confusion matrix is basically a matrix of $2x2$ with predicted values on one side and actual on another [15].

Accuracy: Referred to as the ratio between total number of predictions and number correct predictions [16]. Mathematically:

$$Accuracy = \frac{TRP + TRN}{TRP + FLP + TRN + FLN} \tag{1}$$

where TRP stands for True Positive, TRN stands for True Negative, FLP stands for False Positive, and FLN stands for False Negative.

Precision: Precision is a ratio of true positives and all positives [16]. Mathematically:

$$Precision = \frac{TRP}{TRP + FLP} \tag{2}$$

Recall: Recall basically measure of model identifying true positives [16]. Mathematically:

$$Recall = \frac{TRP}{TRP + FLN} \tag{3}$$

F1-score: It is a measure computed via Harmonic mean between precision and recall [16]. Mathematically:

$$F1\ score = 2 * \frac{PRC * RCL}{PRC + RCL} \tag{4}$$

where PRC: Precision, RCL: Recall

Recall ROC Curves (Receiver Operating Characteristic Curve): This is obtained by plotting true positive rate vs. false positive rate. In this study the models we have used classifies that either a person is COVID positive or not supported the possibilities generated for every class, we are able to decide the edge of the possibilities likewise [16].

Consider that threshold line worth of 0.4. This would imply that the model is capable of classifying the patient's probability of having COVID higher than 0.4. Clearly, this generates a high recall worth and scale back the amount of False Positives.

AUC Interpretation: This implies that for the bottom factor $(0, 0)$, the edge is about at 1.0. Therefore the classifier will categorize all sufferers as now no longer having a coronary heart sickness. At a factor (1, 1), the edge is about at 0.0. This will lead to classify all sufferers as having a coronary sickness [16]. The values of FPR and TPR for the edge values among 0 and 1 provides the relaxation of the curve. We examine that, as TPR for the edge values among 0 and 1 for FPR near zero, we're reaching a TPR of near 1. This is while the version will expect the sufferers having coronary sickness nearly perfectly [16]. This limitations is known as the Area Under Curve i.e., AUC. We it as a metric of an excellent version so that starting from zero to 1, we need to intent for a excessive cost of AUC. Models with a excessive AUC are known as as fashions with desirable skill [16].

For example if we got a value of 0.88 as AUC which is quite good enough and that means our model can differentiate between the positive and negative

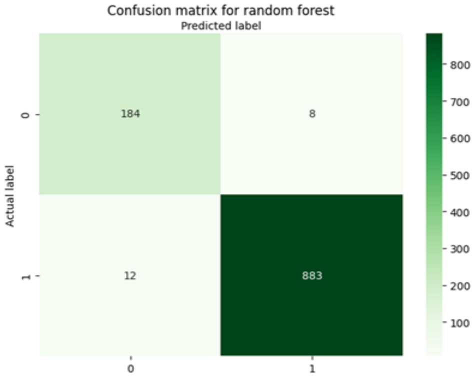

Fig. 4: Random Forest

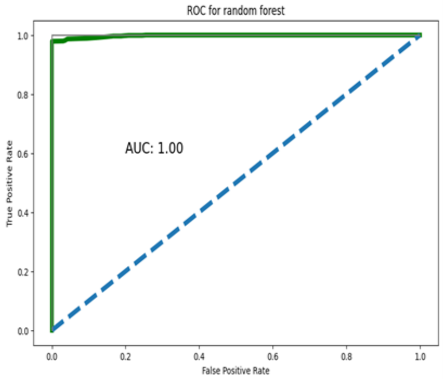

Fig. 5: ROC for Random Forest

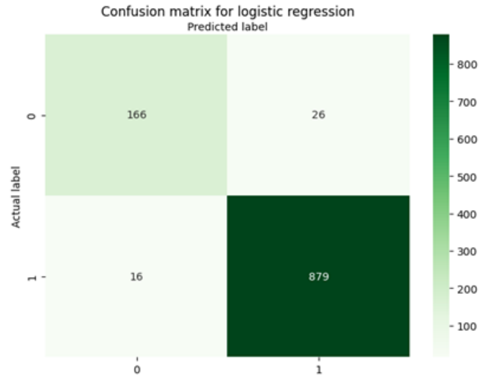

Fig. 6: Logistic Regression

patients approximately 88% of the time [16].

3.6.1 Decision Tree

Confusion matrix of decision tree depends on two labels (Predicted label, Actual Label) and four matrices such as: True Positive (TRP), False positive (FLP), False Negative(FLN) and True Negative (TRN). In this research used five thousand cases to measure accuracy of the model. TRP cases are one hundred sixty six(166), FLP cases are twenty six (26), FLN cases are sixteen (16) and TRN cases are eight hundred and seventy nine (879) shown in Figure 2.

3.6.2 Random Forest

Confusion matrix of random forest depends on two labels (Predicted label, Actual Label) and four matrices such as: contain True Positive (TRP), False positive (FLP), False Negative(FLN) and True Negative (TRN). In this research used five thousand cases to measure accuracy of the model. TRP cases are one hundred eighty four (184), FLP cases are eight (8), FLN cases are twelve (12) and TRN cases are eight hundred and eighty three (883) shown in Figure 4.

3.6.3 Logistic Regression

Confusion matrix of Logistic Regression depends on two labels (Predicted label, Actual Label) and four matrices such as: contain True Positive (TRP), False positive (FLP), False Negative (FLN) and True Negative (TRN). In this research used five thousand cases to measure accuracy of the model. TRP cases are one hundred sixty six(166), FLP cases are twenty six (26), FLN cases are sixteen (16) and TRN cases are eight hundred and seventy nine (879) shown in Figure 6.

3.6.4 Deep Learning

This research used five thousand and four hundred cases to measure accuracy of the model. TRP cases are three hundred and four (304), FLP cases are fourteen (14), FLN twenty (20) and TRN cases are one thousand two hundred and ninety three (1293) as shown in Figure 8. Figure 9 shows the ROC curve for the deep learning model with an area under curve of 1.00. Figure 10 shows the representation of the correlation matrix using a heatmap for the visualization of different features or attributes.

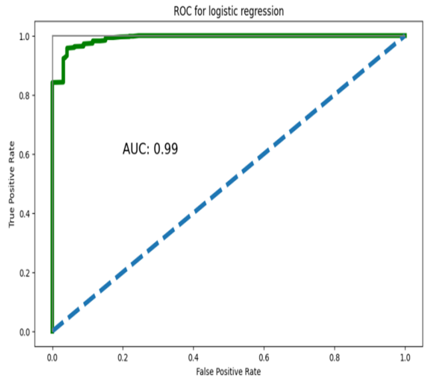

Fig. 7: ROC For Logistic Regression

Finally, in Table 1 , we show the overall accuracy of various ML models considered in this paper. The major features in this table are: Precision, Accuracy and F1-score.

4 Results

4.1 Performance Evaluation

Twenty attributes were considered for the prediction of coronavirus. Machine learning models were developed and applied. These algorithms were trained with a dataset containing 20 different features. Table 2 shows the performance evaluation of the results. We checked the performance of the model by using an 80-20 train-test-split approach [9].

In terms of accuracy, we see that the RF achieves the highest accuracy by reaching up to 0.98, followed

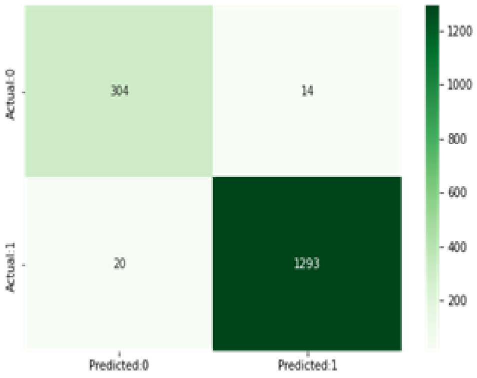

Fig. 8: Deep Learning

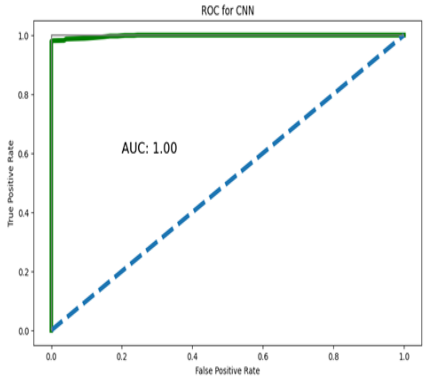

Fig. 9: ROC for CNN

by DT and LR with an accuracy of 0.96. The DL achieved an accuracy of 0.87. While comparing the precision, again the RF outperforms other by reaching a precision of 0.99. DT and LR could reach up to 0.97 whereas DL achieved only 0.88. In case of recall, all the models obtained a value 0.98 except DL which remained on 0.88. For F1, RF again outperform others by going up to 0.98, DT and LR achieved 0.97 and DL obtained 0.88.

TABLE 1: Accuracy of Confusion Matrix of Various ML models

Measure	Value			
	DT	RF	LR	DL
STV	0.9121	0.9388	0.9121	0.9388
SPC	0.9713	0.9910	0.9713	0.9910
PRC	0.8646	0.9583	0.8646	0.9583
NPV	0.9821	0.9866	0.9821	0.9866
FPR	0.0287	0.0090	0.0287	0.0090
FDR	0.1354	0.0417	0.1354	0.0417
FNR	0.0879	0.0612	0.0879	0.0612
ACC	0.9614	0.9816	0.9614	0.9616
F1	0.8877	0.9485	0.8877	0.9485
MCC	0.8648	0.9373	0.8648	0.9373

DT: Decision Trees, RF: Random Forest, LR: Logistic Regression, DL: Deep Learning.
STV: Sensitivity, SPC: Specificity, PRC: Precision, NPV: Negative Predictive Value, FPR: False Positive Rate, FDR: False Discovery Rate, FNR: False Negative Rate, ACC: Accuracy, F1: F1 Score, MCC: Matthews Correlation Coefficient.

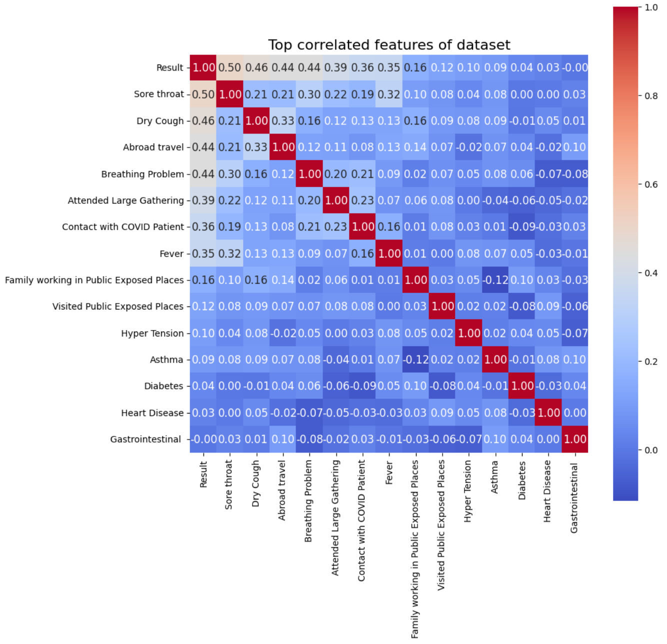

Fig. 10: Representation of correlation matrix of a COVID-19 dataset

TABLE 2: Overall results obtained from the COVID-19 dataset

Models	ACC	PRC	RCL	F1
DT	0.96	0.97	0.98	0.97
RF	0.98	0.99	0.98	0.98
LR	0.96	0.97	0.98	0.97
DL	0.87	0.88	0.88	0.88

DT: Decision Trees, RF: Random Forest, LR: Logistic Regression, DL: Deep Learning.
ACC: Accuracy, PRC: Precision, RCL: Recall, F1: F1 Score

5 Conclusion

This study has been done to find a suitable model for the prediction of Coronavirus. In this research few algorithms like Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Deep Learning (CNN) were selected. After the overall Comparison of results, RF turns out to be the best model. Random Forest has better accuracy than the other three models such as Decision Tree and Logistic Regression. LR and DT turn out to be the second-best algorithm after Random Forest. RF has better results for other features such as precision, recall, and f1-score as compared to LR and DT and deep learning. As we know already a lot of research is going on coronavirus using machine learning to forecast the early detection of coronavirus. This study not only just predicts the COVID-19 based on the symptoms dataset which is used in this study but also finds the best prediction algorithm among the other models which are used in this research. For Future Work, the results of this study can help design an AI prediction tool to help health professionals. Moreover, this work can be helpful to the new researchers who want to work on designing prediction models for disease detection (not limited to just COVID-19 but applies also to other diseases too).

References

[1] John Hopkins Medicine, what is coronavirus?, Online: https://www.hopkinsmedicine.org/health conditions-and-diseases/coronavirus (accessed on 12 Dec 2021).

[2] World Health Organization (WHO), Coronavirus Disease COVID-19, Available online: https://www.who.int/health-topics/coronavirus (accessed on 12 Dec 2021).

[3] Zoabi, Yazeed, Shira Deri-Rozov, and Noam Shomron. “Machine learning-based prediction of COVID-19 diagnosis based on symptoms.” npj digital medicine 4, no. 1 (2021): 1-5.

[4] Manapure, Pranali, Kiran Likhar, and Hemlata Kosare. “Detecting COVID-19 in X-ray images with keras, tensor flow, and deep learning.” assessment 2, no. 3 (2020).

[5] Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. “Coronavirus disease (COVID-19) cases analysis using machine-learning applications.” Applied Nanoscience (2021): 1-13.

[6] Muhammad, L. J., Ebrahem A. Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and Ibrahim Alh Mohammed. “Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset.” SN computer science 2, no. 1 (2021): 1-13.

[7] Chan, Jasper Fuk-Woo, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, Fanfan Xing et al. “A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.” The lancet 395, no. 10223 (2020): 514-523.

[8] Shankar, VirenViraj, Varun Kumar, Umesh Devagade, Vinay Karanth, and K. Rohitaksha. “Heart disease prediction using CNN algorithm.” SN Computer Science 1, no. 3 (2020): 1-8.

[9] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. “Heart disease prediction using machine learning techniques.” SN Computer Science 1, no. 6 (2020): 1-6.

[10] JavaTPoint, “Decision tree classification algorithm”, Available online: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed on 25 Feb 2022)

[11] Krishna, M. Manoj, M. Neelima, Mane Harshali, and M. Venu Gopala Rao. “Image classification using deep learning.” International Journal of Engineering & Technology 7, no. 2.7 (2018): 614-617.

[12] Nassif, Ali Bou, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. “Speech recognition using deep neural networks: A systematic review.” IEEE access 7 (2019): 19143-19165.

[13] Li, Yu, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. ”Deep learning in bioinformatics: Introduction, application, and perspective in the big data era.” Methods 166 (2019): 4-21.

[14] Alakus, Talha Burak, and Ibrahim Turkoglu. “Comparison of deep learning approaches to predict COVID-19 infection.” Chaos, Solitons & Fractals 140 (2020): 110120.

[15] “Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score.” Online: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262. (accessed on 04.11.2022)

[16] “Precision vs. Recall – An Intuitive Guide for Every Machine Learning Person.” Online: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning. (accessed on 04.11.2022)

[17] Saleem, Farrukh, Abdullah Saad Al-Malaise Al-Ghamdi, Madini O. Alassafi, and Saad Abdulla AlGhamdi. “Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review.” International journal of environmental research and public health 19, no. 9 (2022): 5099.

[18] Alyasseri, Zaid Abdi Alkareem, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A. Awadallah, Ammar Kamal Abasi, Sharif Naser Makhadmeh, Osama Ahmad Alomari et al. “Review on COVID-19 diagnosis models based on machine learning and deep learning approaches.” Expert systems 39, no. 3 (2022): e12759.

[19] Mohan, Senthilkumar, Ahed Abugabah, Shubham Kumar Singh, Ali Kashif Bashir, and Louis Sanzogni. “An approach to forecast impact of Covid-19 using supervised machine learning model.” Software: Practice and Experience 52, no. 4 (2022): 824-840.