




IJCRR
Section: Healthcare
ISI Impact Factor
(2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Copyright@IJCRR

Prediction of COVID-19 Possibilities using K-Nearest Neighbour Classification Algorithm

Prasannavenkatesan Theerthagiri*, Jeena Jacob I, Usha Ruby A, Vamsidhar Yendapalli

Department of Computer Science and Engineering, GITAM School of Technology, GITAM University, Bengaluru-561203, India

ABSTRACT

Introduction: COVID-19 is an acute respiratory illness that directly affects the lungs. It is much needed to predict the possibility of occurrence of COVID-19 based on their characteristics.

Objective: This paper studies the different machine learning classification algorithms to predict the COVID-19 recovered and deceased cases.

Methods: The k-fold cross-validation resampling technique is used to validate the prediction model. Aim and The prediction scores of each algorithm are evaluated with performance metrics such as prediction accuracy, precision, recall, mean square error, confusion matrix, and kappa score. For the preprocessed dataset, the k-nearest neighbour (KNN) classification algorithm produces 80.4 % of predication accuracy and 1.5 to 3.3 % of improved accuracy over other algorithms.

Results: The KNN algorithm predicts 92 % (true positive rate) of the deceased cases correctly, with 0.077% of misclassification. Further, the KNN algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. Also, it produces the receiver operator characteristic curve with an output value of 82 %.

Conclusion: Based on the prediction results of various machine learning classification algorithms on the COVID-19 dataset, this paper shows that the KNN algorithm predicts COVID-19 possibilities well for the smaller (730 records) dataset than other algorithms.

Key Words: COVID-19, Prediction, Classification, Machine learning algorithms, KNN

INTRODUCTION

Covid-19 a disease that was caused due to a virus called coronavirus.¹⁻³ It became a global epidemic disease, according to the World Health Organisation (WHO). It was started at the Wuhan of China at the end of 2019. The symptoms of this disease at an early stage are cough, fever, fatigue, and myalgias.¹ Later the patients suffer from heart damages, respiratory problems, and secondary infection situations. Spreading of COVID-19 happens very fast because it spreads through contact, contaminated surfaces, and infected fluids. When the condition of the patient becomes worse with respiratory issues, the patient needs to be treated in an intensive care unit with ventilation.

The mortality of this disease increases day by day, and this disease becomes a big threat to humankind of the entire world. Along with the clinical researches, the analysis of re-

lated data will help mankind. Many types of research have already been done on the Computed Tomography (CT) images of the patients, their symptom-based analysis, and the influencing factors.^{4,5} Researches on CT images were done for identifying the characteristics of the disease and also diagnosing the disease early. CT images of COVID-19 cases have similarities in terms of inward and circular diffusion.⁴

The classifications of Covid-19 are Influenza-A viral pneumonia, Covid-19, and healthy one.⁴ The research is done based on CT images of 618 images with 224 images of Influenza patients, 219 images of COVID-19 patients, and 175 healthy humans, and they achieved 87.6% accuracy. Another study was done for segmenting and quantifying the infection of CT images.⁵ They used the CT images of the chest and lung, and they implemented it using deep learning techniques. They used 249 images for training and 300 images

Corresponding Author:

Prasannavenkatesan Theerthagiri, Department of Computer Science and Engineering, GITAM School of Technology, GITAM University, Bengaluru-561203, India; Email: vprasann@gitam.edu

ISSN: 2231-2196 (Print)

ISSN: 0975-5241 (Online)

Received: 01.11.2020

Revised: 01.01.2021

Accepted: 29.01.2021

Published: 30.03.2021

for testing and achieved an accuracy of 91.6%. Pathological tests and analysis of CT images take some time. So researches are done based on the possibility of disease prediction based on the symptoms. This work uses some classification techniques for predicting the possibility of occurrence of COVID-19 based on their characteristics.

Most of the existing works concentrate on COVID-19 prediction using images. This work proposes the patient data-based prediction of COVID-19 possibilities (recovered or deceased) using the KNN classification algorithm. The prediction performance over 730 records of the COVID-19 patient dataset is evaluated using the KNN algorithm. Further, in this work, the MSE rate, kappa scores, and classification report are analyzed for the proposed KNN algorithm. The results of this research work, suggest that the proposed KNN algorithm produces better results for the smaller dataset than other algorithms.

The organization of this paper is as follows. Section 2 gives the related work of classification techniques. Section 3 discusses the different machine learning models. Section 4 analyses the performance metrics, experimental analyses, and results, and Section 5 gives the concluding remarks with future work.

The emergence of Artificial Intelligence (AI) transformed the world in all fields. Machine learning (ML), a subset of AI helps the human to find solutions for highly complex problems and also plays a vital role in making human life sophisticated. The application areas of ML include business applications, intelligent robots, autonomous vehicle (AV), healthcare, climate modelling, image processing, natural language processing (NLP), and gaming. The learning of ML mimics human intelligence, and it is implemented based on the trial and error method. The instructions to the algorithm were given mainly using control statements such as conditional if.⁶ Many prediction based algorithms are available in ML.⁷ The ML techniques are used for classification and prediction in various fields like disease prediction, stock market, weather forecasting, and business.

In the medical field also, many ML algorithms are used for disease prediction⁸ like coronary artery disease⁹, predicting cardiovascular disease¹⁰, and prediction of breast cancer.¹¹ Several types of research are also done for COVID-19 confirmed case live forecasting¹² and for predicting the COVID-19 outbreak.¹³ These works will aid the higher authorities of the country in taking decisions to handle the situation by foreseeing.¹⁴ At first, the COVID-19 was misinterpreted as pneumonia.¹⁵ But the failure of multi-organs and high mortality rates made it a pandemic in the whole world.¹⁶

Classification techniques are broadly categorized into semi-supervised,^{17,18} supervised¹⁹ and unsupervised.²⁰⁻²³ Supervised learning takes information about the classes and learns

based on that information. Based on this knowledge, this technique can predict the classes for new data. In unsupervised learning, the information about the classes is unknown. The clustering of similar data is done by identifying the similarity among themselves. Semi-supervised techniques know some data information, and the classification is done based on it. Logistic Regression is used for relationship analysis between various dependent variables.^{24,25} Basically, it was used for identifying the existence of a class or event. This was further extended to classify more objects. Artificial neural network (ANN) is based on learning and classifies effectively.²⁶⁻³¹ Here, the nodes are arranged in the input layer, hidden layer, and output layer. Based on the objective function and the number of hidden layers will vary. Support Vector Machine (SVM) is another classification technique that separates the variables using a hyperplane.³²⁻³⁵

Many classification and prediction algorithms are applied to study the possibility of spreading COVID-19. The research was done on the occurrence of asymptomatic infection, and they found it is higher (15.8%) in children under 10 years.³⁶ Some studies have done in identifying the symptoms and identified having lesser senses of taste and smell are the signs of Covid-19.³⁷ Another work also studied the transmission process of this disease.³⁸

In recent years, predictive medical analysis using machine learning techniques has tremendous growth with promising results. The machine learning algorithms are effectively applied in numerous types of applications in diverse fields. Many kinds of research have proved that the machine learning predictive algorithms had provided better assistance for clinical supports as well as for decision making based on the patient data.³⁹ In the healthcare field, disease predictive analysis is one of the useful and supportive applications of machine learning prediction algorithms. This research work applies the predictive disease analysis using KNN machine learning prediction algorithms for the novel COVID-19 disease.

The contribution of the proposed work is listed as follows:

- This research work investigates COVID-19 patient data to assess the outcome possibilities of the patient.
- The KNN classification algorithm is proposed in this work to predict the outcome possibilities of patients such as recovered or deceased.
- The prediction results of the proposed KNN algorithm is evaluated for the accuracy rate, mean square error rate, Kappa score, the area under the curve, indices, sensitivity, specificity, and f1 score values.
- This work considers only two parameters of the patients with 730 records, and KNN algorithm based outcome prediction results are compared to other algorithms.

MATERIALS AND METHODS

Data Preprocessing and Cleaning

The COVID-19 dataset from the Kaggle is taken for the predictive analysis in this research work.⁴⁰ The considered dataset was cleaned using the data preprocessing and data cleaning methodologies, then the resulted dataset has been considered for several numbers of experiments over different classification algorithms. The COVID-19 dataset contains the patient's details with recovered and deceased status. The vital patient's information is used to diagnose and predict the COVID-19 disease among the infected population.

The considered COVID-19 dataset contains 100284 records. The dataset contains features of patients such as patient number, state patient number, date announced, estimated onset date, age bracket, gender, detected city, detected district, detected state, state code, current status, notes, contracted from which patient (suspected), nationality, type of transmission, status change date, source_1, source_2, source_3 (source of patient information), backup notes, num cases, entry_id.⁴⁰

The data preprocessing and cleaning process (data imputation-mean technique) removes the missing and outlier data values from the dataset. The resulted dataset after preprocessing is reduced to 730 records with three required relevant features of patient details. In the dataset, there are 730 patient details, out of which 156 cases are in the class of 'recovered from COVID-

at the time of infection by the COVID-19 virus; 2. Gender- classifies whether the patient is male or female; 3. Outcome-denotes whether the patient has been recovered from COVID-19 disease or deceased due to COVID-19 disease. Figure 1(a) illustrates the population infected by COVID-19 concerning age. Figure 1(b) and Figure 1(c) depict the count plot of gender and outcome of COVID-19, respectively.

This research work analysis the prediction of recovered and deceased patients infected by COVID-19. Different classification models are applied to the COVID-19 dataset, and its performance in terms of accuracy, error rates, etc. are evaluated. The classifiers evaluated in this research work are Logistic Regression (LR), K-Nearest Neighbors Classifier (KNN), Decision Tree (DT), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP).

Logistic Regression (LR)

One of the simple and powerful prediction algorithms is logistic Regression. The logistic Regression uses the sigmoid function for predictive modelling of the given problem. It models the dataset maps them into a value between 0 and 1. The logistic Regression performs the predictive analysis based on the relationship between the binary dependent variable and the other one or more independent variables from the given dataset. To predict the output value (Y), the input values ($X1, X2, \dots, Xn$) are linearly combined using the coefficient values.⁴¹ Let us consider, 'Y' as the output prediction variable and $X1$ and $X2$ are input variables, then the logistic regression equation is given as (1),

$$Y = \frac{1}{2} \left[\frac{e^{(mX1+c)}}{1 + e^{(mX1+c)}} + \frac{e^{(mX2+c)}}{1 + e^{(mX2+c)}} \right] \quad (1)$$

Where 'c' represents the intercept, 'm' is the coefficient of input value $X1$ and $X2$ (in our case, $X1, X2$ are age, gender). The coefficient value 'm' can learn from the training dataset for each input value ($X1, X2$).⁴¹ This work is to classify the deceased and recovered cases of the COVID-19 disease using the equation (1).

K-Nearest Neighbors (KNN) Classifier

K-Nearest Neighbors algorithm is the non-parametric algorithm. The learning and prediction analysis is performed based on the given problem or dataset. The KNN classification model, the prediction is purely based on neighbor data values without any assumption on the dataset. In KNN, 'K' represents the number of nearest neighbor data values. Based on 'K', i.e., the number of nearest neighbors, the decision is made by the KNN algorithm on classifying the given dataset.⁴¹ The KNN model directly classifies the training dataset. It means the prediction of a new instance is made by searching the similar 'K' neighbour instances in the entire training set and classifying based on the class of highest instances. A

Table 1: Sample record of cleaned dataset

Age	Gender	Outcome
13	Female	Recovered
96	Male	Recovered
89	Female	Recovered
85	Male	Recovered
27	Male	Recovered
69	Female	Deceased
26	Male	Recovered
65	Male	Deceased
76	Male	Deceased
45	Female	Recovered

19 disease' and 574 cases are in the class of 'deceased by the COVID-19 disease' with 99554 records are missing required essential values. Two numerical features from the dataset are taken as the input attributes, and one feature is considered as the output attribute. The COVID-19 patient's information is presented in Table 1.

The patient features such as age and gender is considered as input variables, and the outcome is taken as the output variable—the features such as 1. Age- denotes the patient

similar instance is determined using the Euclidean distance formula. Euclidean distance is the square root of the sum of squared differences between the new instance (x_j) and the existing instance (x_i).⁴²

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^n (x_{ik} - y_{jk})^2} \quad (2)$$

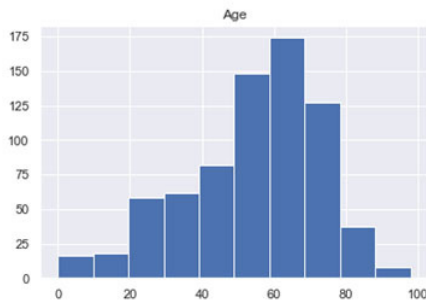


Figure 1(a) Population vs Age.

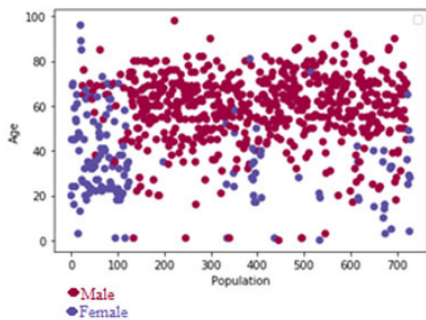


Figure 1(b) Gender.

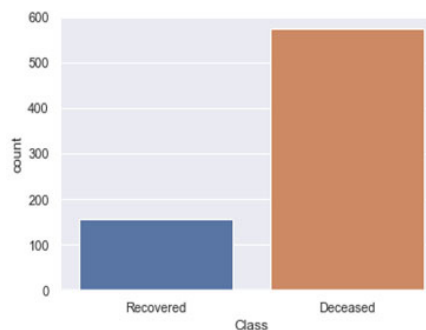


Figure 1(c) Outcome.

Decision Tree (DT)

The decision tree algorithms are the powerful prediction model used for both classification and regression problems. The decision tree models are represented in the form of a binary tree. It means the given problem/dataset is solved by splitting or classifying them as a binary tree. In the decision tree, the prediction is made by taking the root node of the binary tree with a single input variable (x), splitting the dataset based on the variable, and its leaf nodes of the binary tree

have resulted as the output variable (y). That is, from the root node, the tree is traversed through each branch with their divisions, and prediction is made based on the leaf nodes. It uses the greedy method for splitting the dataset in a binary manner.⁴³

In this research work, the COVID-19 dataset with two inputs (x) is taken as age, gender, and output is whether the patient is recovered or deceased. The decision tree classification algorithm uses the Gini index function to determine the impurity level of the leaf nodes for the predictions. The Gini index function (G) is given in equation (3).

$$G = \sum_{i=1}^n xk(1-xk) \quad (3)$$

Where ' x ' is the proportion of training instances in the input class ' k '. Binary tree representation of the dataset predicts straightforward.⁴³

Support Vector Machines (SVM)

The support vector machine can handle categorical and continuous variables. Also, the SVM model works well on classification and regression problems. The support vector machine is a classification algorithm that creates the hyperplanes for each class labels in the multidimensional space by employing the margin values. The SVM intends to maximize the margins among different classes by optimally separating hyperplanes.⁴⁴ The hyperplane is a data instance of the given dataset used by the support vectors. The margin is the maximum distance between the support vector and the hyperplane.⁴⁴ If the given dataset is linear bounded, then linear SVM can be adopted, and the dataset is non-linear bounded, then Non-linear SVM can be adopted for the classification tasks.⁴⁵

Let us consider a dataset $(A_1, B_1, \dots, A_n, B_n)$; where (A_1, \dots, A_n) is the set of the input variable, (B_1, \dots, B_n) is the output variable, and ' C ' is the intercept, then the SVM classifier⁴⁴ is given as like equation (4).

$$SVM = \sum_{i=1}^n \beta_i - \frac{1}{2} \sum_{i,j=1}^n b_i b_j C(a_i, a_j) \beta_i \beta_j \quad (4)$$

In the equation (4), $i=1, 2, \dots, n$; and $C = b_i \beta_i + b_j \beta_j$. The SVM equation (4), is used in this research work to classify the deceased and recovered cases of the COVID-19 disease.

Multilayer Perceptron (MLP)

The Multilayer Perceptron algorithm is suitable for classification problems and predictive analysis. The MLP is the classical neural network with one or more layers of hidden neurons. It comprises the input layer (where the data variables are fed), the hidden layer (with function to operate on the data), and the output layer (contains the predicted val-

ues). MLP uses the back-propagation to learn from the given input and output dataset. The activation function $A_j(X, W)$ of the MLP is the summation of the inputs (X) multiplied with respective weights (W_{ij}) as represented in equation (5). The output function (O_j) with the sigmoid activation function of the MLP back-propagation⁴⁶ algorithm is given in equation (6).

$$A_j(X, W) = \sum_{i=0}^n (X_i, W_{ij}) \quad (5)$$

$$O_j(X, W) = \frac{1}{1 + e^{-A_j(X, W)}} \quad (6)$$

RESULTS AND DISCUSSION

This section summarizes the prediction results of the logistic Regression, k-neighbours classifier, decision tree, support vector machines, and multilayer perceptron algorithms.

Cross-Validation

To evaluate and validate the performance of the machine learning model, resampling methods are adopted. This method estimates the prediction ability of the machine learning algorithm on new unseen input data. The k-fold cross-validation is one of the resampling procedure used in this work to validate the machine learning models on the limited data sample. The 'k' represents the number of times the data model is to split. Each split of the data sample is called a subsample or sampling group. These subsamples are used to validate the training dataset. In this work, the 'k' value is chosen as 7. Therefore, it can be called a 7-fold cross-validation resampling method. The 7-fold cross-validation method intends to reduce the bias of the prediction model.⁴⁷

Performance Metrics

Typically, the performance of the machine learning prediction algorithms measured by using some metrics based on the classification algorithm. In this work, the prediction results are evaluated by using the metrics such as accuracy, mean square error (MSE), root means square error (RMSE), Kappa score, confusion matrix, the area under the curve (ROC_AUC), classification performance indices, sensitivity, specificity, and f1 score values.

Mean Square Error (MSE): It is the average of the squared difference between predicted results (P_i) and actual results (A_i). It is calculated by using the equation is given in (7), where n is the number of samples.⁴⁸

$$MSE = \frac{1}{n} \sum_{i=1}^n (P_i - A_i)^2 \quad (7)$$

Root Mean Squared Error (RMSE): The RMSE is the square root of the average of squared differences between predicted and actual results, likewise given in (8). It depicts the inconsistencies among the observed and predicted values.⁴⁹

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - A_i)^2} \quad (8)$$

Accuracy: The accuracy of the prediction algorithm is the ratio of the total number of correct predictions of class to the actual class of the dataset. Equation (9) calculates the accuracy of the model. Typically, any prediction model produces four different results, such as true positive (TP), true negative (TN), false positive (FP), and false-negative (FN).⁴²

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (9)$$

Precision: the precision of the prediction algorithm is the number of correctly predicted recovered COVID-19 cases that is belonging to the actual recovered COVID-19 cases.^{42,47}

$$Precision = \frac{TP}{TP + FP} = \frac{\text{True Positive}}{\text{Total predicted positive}}$$

Recall: recall of the prediction algorithm is the number of correctly predicted recovered COVID-19 cases made out of all recovered COVID-19 cases in the dataset. It is a true positive rate.^{42,47}

$$Recall = \frac{TP}{TP + FN} = \frac{\text{True Positive}}{\text{Total predicted positive}}$$

F1 Score: it is the measure of the balanced score (harmonic mean) of both precision and recall.⁴²

$$F1Score = \frac{Precision \times Recall}{Precision + Recall}$$

Cohen's kappa Score: Cohen's kappa score estimates the consistency of the prediction model. It compares the result of the predicted model with actual results. It is a statistic value between 0 and 1. A value near 1 might have the great consistency.⁴⁷

$$K = \frac{[TP + TN / N] - [(TP + FN)(TP + FP)(TN + FN) / N^2]}{1 - [(TP + FN)(TP + FP)(TN + FN) / N^2]}$$

Confusion matrix: The confusion matrix provides a complete insight into the performance of a prediction model. It produces prediction results in the matrix form with the information of the number of correctly predicted cases, incorrectly predicted cases, errors of incorrect, and correct prediction cases.⁴⁷

Receiver Operating Characteristic (ROC)-Area Under Curve (ROC_AUC): The ROC_AUC curve is a graphical illustration of the performance of the prediction model.⁴⁷ The ROC curve is the relationship between the recall and precision over varying threshold values. The threshold is the positive predictions of the model. The ROC_AUC curve plotted by keeping the x-axis a false positive rate and the y-axis as a true positive rate. Its value ranges from 0 to 1.⁴⁷

Performance Evaluation

In most of the research works, the accuracy of the prediction model has been taken as one of the common performance metrics while working on a prediction algorithm.⁴² In this work, the prediction accuracy (that is, whether the COVID-19 infected patient is recovered or deceased) of different machine algorithms (logistic Regression, k- nearest neighbour, decision tree, support vector machines, and multilayer perceptron) are determined. Each classification model has a different prediction accuracy based on its hyperparameters and a certain level of improvement over other prediction models. This work considers 70 % dataset for training and 30 % of the data samples for testing in classification algorithms. In this work, each model's accuracy is compared, and its prediction results are summarized in Table 2.

Table 2: Accuracy score of classifiers

S. No	Classifier	Accuracy	Kappa
	Logistic Regression (LR)	78.5388	0.4109
	K Neighbors Classifier (KNN)	80.3653	0.469
	Decision Tree (DT)	75.3425	0.3043
	Support Vector Machines (SVM)	78.9954	0.4266
	Multi-Layer Perceptron (MLP)	77.1689	0.03411

In Table 2, the classification algorithms such as logistic Regression, k-nearest neighbour, decision tree, support vector machines, and multilayer perceptron have the prediction accuracy of 78.5388, 80.3653, 75.3425, 78.9954, and 77.1689 respectively. Whereas, the k-nearest neighbour algorithm predicts the outcome of the COVID-19 cases (based on age and gender) more accurately than the other algorithms. Here, the dataset was tested with several 'k' values and the 'k' value 2, which classifies the KNN algorithm into two clusters as recovered and deceased with reduced errors. It works by calculating the distance between the test data and training data. For each data points based on the distance values, the testing datasets are classified. Such that the KNN algorithm produces a higher classification rate than the other algorithms.

Cohen's kappa score for the KNN algorithm is also high than other algorithms. Cohen's kappa score estimates the consistency of the classification algorithm based on its predictions. Figure 2 depicts the accuracy scores of different classifica-

tion algorithms. From Figure 2, we can see that the k-nearest neighbour algorithm has the highest accuracy of 80.4. The KNN algorithm has 1.5 to 3.3 % of improved accuracy as compared to LR, DT, SVM, and MLP algorithms. The KNN algorithm works by classifying the data point of the COVID-19 dataset based on the similarity. The closely matching data point are grouped. Thus, it increases the accuracy rate of the KNN algorithm.

Table 3 presents the performance error metrics of the various machine learning algorithms. The error metrics mean square error and root mean square error values for each algorithm is evaluated.

S. No.	Classifier	MSE	RMSE
	Logistic Regression (LR)	0.2146	0.4633
	K Neighbors Classifier (KNN)	0.1963	0.4431
	Decision Tree (DT)	0.2466	0.4966
	Support Vector Machines (SVM)	0.21	0.4583
	Multi-Layer Perceptron (MLP)	0.2283	0.4778

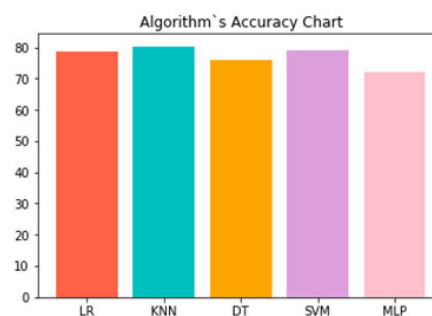


Figure 2: Prediction of accuracy.

The logistic Regression, k-nearest neighbour, decision tree, support vector machines, and multilayer perceptron have the MSE error rate as 0.2146, 0.1963, 0.2466, 0.21, and 0.2283, respectively. As per Figure 3 (a), the KNN classification algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. The KNN algorithm classifies the testing dataset by calculating the Euclidean distance between the new (testing) instance (x_j) and the existing (training) instance (y_j). Therefore, it results in lower error rates.

Similarly, the KNN's RMSE error rate also very low (0.44) as compared to the error rates of LR (0.46), DT (0.50), SVM (0.45), and MLP algorithms. As depicted in Figure 3(b), the KNN classification algorithm produces the highest consistency among the evaluated algorithms as 0.47. The SVM algorithm offers the next highest consistency (0.42) on correctly predicting the COVID-19 cases. Moreover, the prediction of the decision tree algorithm has the lowest consistency value as 0.30. The LR and MLP have consistency values of 0.41 and 0.34, respectively.

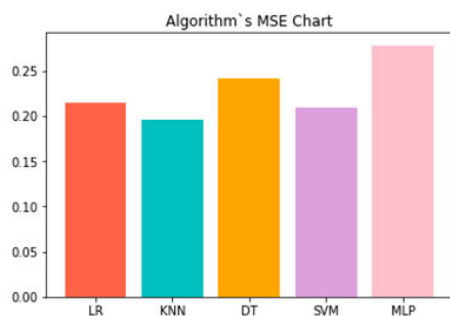


Figure 3(a): MSE rates.

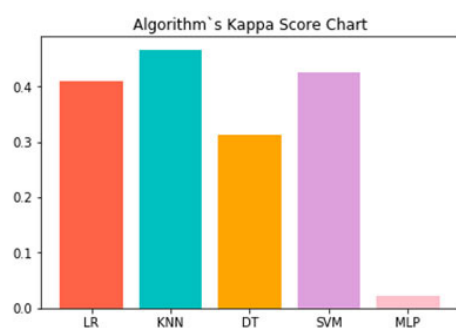


Figure 3(b): Cohen's kappa scores.

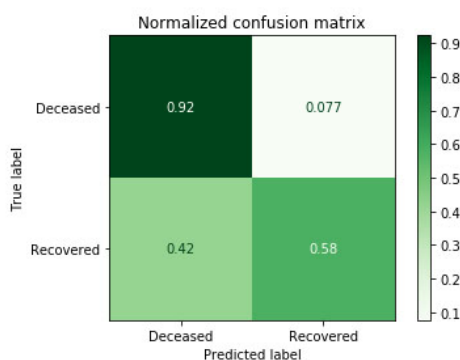


Figure 4(a): Normalized Confusion matrix.

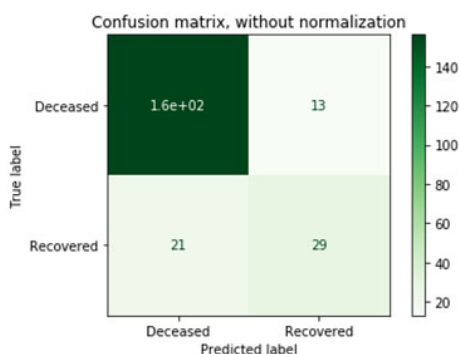


Figure 4(b): Confusion matrix (no normalization).

Figure 4(a) illustrates the normalized confusion matrix of the

k-nearest neighbour classification algorithm. In all classification algorithm, 30 % of the data samples are taken for testing with the 70 % training dataset. In Figure 4(a), the x-axis represents the percentage of predicted values and the y-axis represents the percentage of true values. It can be seen that the KNN algorithm predicts 92 % (true positive) of the deceased cases correctly, with 0.077 % (false positive) of misclassification. Similarly, in Figure 4(b) the confusion matrix without normalization is depicted, where 160 cases are correctly predicted as deceased cases, and 13 cases are misclassified. Further, 31 cases are correctly predicted as deceased cases, and 31 cases are misclassified. Also, it correctly predicts 29 patients (true negative) as the recovered cases, and 21 cases are misclassified (false negative).

Figure 5 is the pictorial representation between the false positive rate and true positive rate in the form ROC area under the curve. The k-nearest neighbour classification algorithm produces the highest value of 0.89 as compared with LR, DT, SVM, and MLP algorithms.

Figure 6 summarizes the performance metrics such as precision, recall, and confusion matrix of the k-nearest neighbour classification algorithm. The KNN algorithm produces the precision (true positive rate) value of 0.82 for the recovered cases and 0.72 for the deceased cases. The recall values for the recovered and deceased cases are 0.92 and 0.50, respectively. Further, the F1 score for recovered and deceased cases is 0.87 and 0.59, respectively.

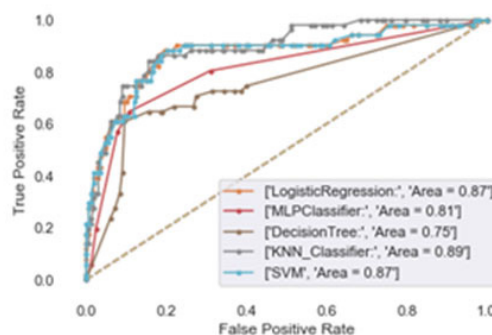


Figure 5: ROC_AUC Curve.

```
KNeighborsClassifier:
-----
MSE: 0.1963470319634703
RMSE: 0.4431106317427628
Kappa_score: 0.46685161071165715
Accuracy: 80.36529680365297

Classification Report:

```

	precision	recall	f1-score	support
Recovered	0.82	0.92	0.87	157
Deceased	0.72	0.50	0.59	62
accuracy			0.80	219
macro avg	0.77	0.71	0.73	219
weighted avg	0.79	0.80	0.79	219

Figure 6: Summary of Performance metrics scores of KNN algorithm.

CONCLUSION AND FUTURE ENHANCEMENTS

Predictive disease analysis is a major application area. This work has implemented Logistic Regression, k-nearest neighbour, decision tree, support vector machines, and multilayer perceptron to classify the COVID-19 dataset. The KNN classification algorithm has 1.5 to 3.3 % of improved accuracy over other machine learning algorithms reported in work. Moreover, the KNN classification algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. To improve the accuracy of predictions, future work will concentrate on predicting the COVID-19 cases using a classification and optimization algorithm

Conflict of interest and Financial support: Nil

Author Contribution:

Prasannavenkatesan Theerthagiri: Conceived and designed the analysis, Collected the data;

Jeena Jacob I: Contributed data or analysis tools;

Vamsidhar Yendapalli: Performed the analysis;

Usha Ruby A: Wrote the paper.

ACKNOWLEDGEMENT

The authors acknowledge the immense help received from the scholars whose articles are cited and included in references to this manuscript. The authors are also grateful to authors/editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

REFERENCES

- Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395(10223):497-506.
- Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, Hussien A, Rathmell J, Thomas B, Chen C, Hales R. Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. *Radiology* 2018;286(1):286-295.
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-118.
- Xie X, Li X, Wan S, Gong Y. Mining X-ray images of SARS patients. *Data Mining: Theory, Methodology, Techniques, and Applications*. *Nature* 2006;23:282-94.
- Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shi Y. Lung infection quantification of COVID-19 in CT images with deep learning. *Lancet* 2020;200304655.
- Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS One* 2018;13(3):e0194889.
- Bontempi G, Taieb SB, Le Borgne YA. Machine learning strategies for time series forecasting. *European business intelligence summer school*. *PloS One* 2012;15:62-77.
- Harrell Jr FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Can Treat Rept* 1985;69(10):1071-1077.
- Lapuerta P, Azen SP, Labree L. Use of neural networks in predicting the risk of coronary artery disease. *Comput Biomed Res* 1995;28(1):38-52.
- Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121(1):293-298.
- Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Proc Comput Sci* 2016;83:1064-1069.
- Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *J Am Med Assoc* 2020;323(16):1545-1546.
- World Health Organization, the World Health Organization. Naming the coronavirus disease (COVID-19) and the virus that causes it. Available: [https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)
- Novel CP. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *Am Heart J* 2020;41(2):145.
- VanHoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC. Identification of a new human coronavirus. *Nature Med* 2004;10(4):368-373.
- Van Der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, et al. Identification of a new human coronavirus. *Nature Med* 2004;10(4):368-373.
- Gómez-Chova L, Camps-Valls G, Bruzzone L, Calpe-Maravilla J. Mean map kernel methods for semisupervised cloud classification. *Trans Geosci Rem Sens* 2009;48(1):207-220.
- Bishop C. Improving the generalization properties of radial basis function neural networks. *Neural Comput* 1991;3(4):579-588.
- Lee JS, Grunes MR, Ainsworth TL, Du LJ, Schuler DL, Cloud SR. Unsupervised classification using polarimetric decomposition and the complex Wishart classifier. *Trans Geosci Rem Sens* 1999;37(5):2249-2258.
- Pahikkala T, Airola A, Gieseke F, Kramer O. Unsupervised multi-class regularized least-squares classification. In 2012 IEEE 12th International Conference on Data Mining 2012 Dec 10:585-594.
- Hang J, Zhang J, Cheng M. Application of multi-class fuzzy support vector machine classifier for fault diagnosis of the wind turbine. *Fuzzy Sets Syst* 2016;297:128-140.
- Kim KI, Jin CH, Lee YK, Kim KD, Ryu KH. Forecasting wind power generation patterns based on SOM clustering. In 2011 3rd International Conference on Awareness Science and Technology (iCAST) 2011 Sep 27 (pp. 508-511).
- Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *J Am Med Assoc* 2016;316(5):533-534.
- Tjur T. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *Am Stat* 2009;63(4):366-372.
- Lee SJ, Hou CL. An ART-based construction of RBF networks. *Neu Net* 2002;13(6):1308-1321.
- Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Cont Sig Syst* 1989;2(4):303-314.
- Durbin R, Rumelhart DE. Product units: A computationally powerful and biologically plausible extension to backpropagation networks. *Neu Comp* 1989;1(1):133-142.

28. Buchla O, Klimek M, Sick B. Evolutionary optimization of radial basis function classifiers for data mining applications. *Trans Syst Man Cyb Part B (Cybernetics)* 2005;35(5):928-947.
29. Yao X. Evolving artificial neural networks. *Neu Comp* 1999 Sep;87(9):1423-1447.
30. Gutiérrez PA, López-Granados F, Peña-Barragán JM, Jurado-Expósito M, Gómez-Casero MT, Hervás-Martínez C. Mapping sunflower yield as affected by *Ridolfia segetum* patches and elevation by applying evolutionary product unit neural networks to remote sensed data. *Compt Electr Agri* 2008;60(2):122-132.
31. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* 1992 Jul 1:144-152.
32. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20(3):273-297.
33. Salcedo-Sanz S, Rojo-Álvarez JL, Martínez-Ramón M, Camps-Valls G. Support vector machines in engineering: an overview. *Mini Know Disc* 2014;4(3):234-267.
34. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. *Neu Net* 2002;13(2):415-425.
35. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y, Wu C. SARS-CoV-2 infection in children. *New Eng J Med* 2020;382(17):1663-1665.
36. Russell B, Moss C, Rigg A, Hopkins C, Papa S, Van Hemelrijck M. Anosmia and ageusia are emerging as symptoms in patients with COVID-19: What does the current evidence say. *New Eng J Med* 2020;14.
37. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, Wang D, et al. Propagation analysis and prediction of the COVID-19. *Infect Dis Mod* 2020;5:282-292.
38. Chetty N, Vaisla KS, Patil N. An improved method for disease prediction using fuzzy approach. In *2015 Second International Conference on Advances in Computing and Communication Engineering* 2015 May 1:568-572.
39. COVID-19 Dataset. Retrieved from: <https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset>
40. Chiang WY, Zhang D, Zhou L. Predicting and explaining patronage behaviour toward web and traditional stores using neural networks: a comparative analysis with logistic regression. *Dec Supp Syst* 2006;41(2):514-531.
41. Altay O, Ulas M. Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbour in children. In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)* 2018 Mar 22 (pp. 1-4). IEEE.
42. Elson J, Tailor A, Banerjee S, Salim R, Hillaby K, Jurkovic D. Expectant management of tubal ectopic pregnancy: prediction of successful outcome using decision tree analysis. *Ultrasound Obstet Gynecol* 2004;23(6):552-556.
43. Schölkopf B, Smola AJ, Bach F. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press; 2002.
44. Kumar GR, Ramachandra GA, Nagamani K. An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets. *Int J* 2014;4(2):272-277.
45. Pal A, Singh JP, Dutta P. Path length prediction in MANET under AODV routing: Comparative analysis of ARIMA and MLP model. *Egy Infor J* 2015;16(1):103-111.
46. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Inform Med Unloc* 2018;10:100-107.
47. Theerthagiri P. FUCEM: futuristic cooperation evaluation model using Markov process for evaluating node reliability and link stability in mobile ad hoc network. *Inform Med Unloc* 2020;26(6):4173-4188.
48. Prasannavenkatesan T. CoFEE: Context-aware futuristic energy estimation model for sensor nodes using Markov model and autoregression. *Int J Commun Syst* 2019:e4248.