FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 4, August, 2023, pp 14 - 26
DOI: https://doi.org/10.33003/fjs-2023-0704-1901

MACHINE LEARNING TECHNIQUES FOR PREDICTION OF COVID-19 IN POTENTIAL PATIENTS

Oyeranmi Adigun, Mohammed Mutiu Rufai, *Okikiola Folasade M., Sunday Olukumoro

Department of Computer Science, Yaba College of Technology, Lagos, Nigeria

*Corresponding authors' email: sade.mercy@yahoo.com

ABSTRACT

The coronavirus pandemic overwhelmed many countries and a shortage of testing kits and centers for exposed patients worsens the situation in most countries. These have prompted the need to quickly predict COVID-19 in patients and stop the spread of the virus. In this research, we present a method for predicting COVID-19 based on symptoms, and to make this system efficient, the dataset was obtained from Afriglobal Laboratory Nigeria, and preprocessing and feature extraction were done on the dataset. Three classifiers, logistic regression, support vector machines, and hybridization of the logistic regression and support vector machines were used to train the data. The test data were evaluated against the model, and the research found that the performance analysis values for accuracy, precision, recall, and F1score for logistic regression (LR) are 91%, 91%, 95%, and 93%, for Support Vector Machines (SVM), 94%, 93%, 100%, and 96% and for the Hybridized model (LR+SVM) are 95%, 94%, 98%, and 96%. To get the parameters needed for the performance evaluation of the classifiers, the confusion matrix method was employed. In comparison to existing methods and studies, the hybridized system performs better than LR and SVM models. As a result, the hybridized model can accurately predict Covid-19.

Keywords: Index Symptoms, coronavirus, Machine learning, confusion matrix, features

INTRODUCTION

COVID-19 is a contagious infection caused by the coronavirus SARSCOV-2 and was discovered in Wuhan, China, in 2019 and is now considered an epidemic (https://www.ncbi.nlm.nih.gov/books/). It is a contagious disease that spread from one individual to another via cough, conversation, and sneezing drops exiting from an infected individual or through touching a virus-infected surface and then touching the mouth and eyes (Sharma et al, 2020, Ashkan et al., 2021). The given signs and symptoms of the infected case are analogous to those of the common flu, which has the potential to lead to lethal conditions such as severe acute respiratory pattern (SARS). As a fully unique infection, there are still some important effects that the medical profession does not realize are contagions (Adigun *et al*., 2020, Hassanien *et al.,* 2020).

So far, millions of people have been infected worldwide; some have recovered, while others have not, resulting in death. Initially, there was no specific vaccine or medicine for COVID-19 treatment. Some important precautions, such as isolation and quarantine, have been the first treatment to be done and it has prevented some complications and deaths (Simionatito et al., 2020). The epidemic has affected every aspect of human life, resulting in an unknown global profit downturn. At the extreme of the epidemic, nearly all governments around the world of each nation declared lockdown. This was the action taken to reduce the spread of the contagion and contain it. It was unpleasant for billions of people all over the world, but they had no choice but to stay indoors and exercise physical distancing and turn to technology to adjust to the new normal.

The lockdown and physical distancing worked well in containing the spread of the contagion, but there is a price to pay. The continuation of the lockdown means keeping numerous people out of jobs and leaving them with little or no income to live for themselves and their families. The international suffering as it did now no longer really motivated the worldwide recession as it did in 2008. (Anshuman *et al*., 2019). As a result of these developments, we have two options, to continue with the lockdown and starve to death or to open up the frugality and face the epidemic. Because of the negative effects on society, no nation will continue to maintain the lockdown, However, the nation's reopening will be critically dependent on mass testing for the contagion, tracing and treating people who have been particularly infected (Lalmuanawna *et al*., 2020). At the height of the global lockdown and in order to maintain physical distancing, people had no other option than to resort to digital technology in all their day-to-day conduct. Technology stood within the gap and played an important part. In our face-to-face commerce, humans have become inextricably linked to smartphones, tablets, computers, and boxes. Apps offering videotape meetings like Zoom, Skype, Google meet, WhatsApp, and Facebook became sources of our relations. The epidemic has brought a turning point that will speed up the digital revolution such as artificial intelligence, data science, and machine learning (Jackins *et al*.,2021).

(Mohanasundaram, A et. al.,2022) During the course of the epidemic, artificial intelligence and machine learning knowledge have been used to develop different algorithms that seek to identify early-stage people who are likely to be infected. These methods make predictions based on the patient personal data, medical symptoms, and also history about discharge time, and tracing contacts of, just recovered patients presented a study for the early ID of patients who can develop severe COVID-19 symptoms. Artificial intelligence recognizes patterns and trends in various disease transmission models which are used in detecting outbreaks, public monitoring, epidemic discovery, and patient contact tracking. (Muhammad et al,2021). They are also important in the fight against COVID-19 because they aid in the diagnosis of the viral epidemic and predict the severity of the contagion.

In the pharmaceutical industry, these models were used to study the genetics and mutations of COVID-19 in order to improve drug prediction and vaccination (Ansary *et al.,* 2020). Medical image recognition using artificial intelligence can be efficient and effective. Due to the rapid advancement of modern algorithms, big data, and hardware computing capability (Bhattacharyya et al., 2022), Convolutional Neural Networks (CNN) have been shown to be an effective medical image recognition technology. Some researchers employ image normalization, image resizing, deep learning, and transfer learning, to categorise chest X-ray images (Mei-Ling Huang, and Yu-Chieh Liao, 2022).

The widespread of the new contagion called coronavirus and its resulting effect has caught everyone around the world off-guard. The virus has completely spread, but it is clear that the crisis is truly global in scope, with scientists and researchers on the front lines fighting it. This includes medical practitioners attempting to heal the sick and the infected while also reducing the mortality rate at the expense of their own health, as well as public health officials tracking the virus and vigilantly implementing measures such as social distancing, and laboratory experts' working to develop drugs, treatments, and vaccines to combat its spread. Coronavirus is a newly identified contagion there is no known pre-existing immunity to fight it in humans. Based on the scientific characteristics and the stories of increasing coronavirus-infected patients throughout the world so far, everyone is assumed to be vulnerable and prone to being infected (Sharma *et al.,* 2020).

Since the outbreak of the contagion, there have been various attempts to understand the virus better and also to curb its spread, numerous model predictions on the COVID-19 epidemic have been reported. The nature of covid-19 has made it very difficult to track and predict, this gives room for its rapid spread. It is very clear that the percentage rates at which humans can predict the deadly disease is very low. Thus, this research work is to aid the early detection and prediction of the virus in potential patients reducing the spread of the deadly contagion.

Related Work

Machine learning plays a serious part in medical prediction and analysis. Though COVID-19 may be a new medical case still works have been done toward COVID-19 prediction using machine learning. Continous research work goes on in this field of exploration. Early detection and diagnosis using artificial intelligence techniques aid in the prevention and control of the COVID-19 pandemic by utilizing various data sources such as CT scans, X-rays, clinical data, and blood sample data. To analyze and predict Covid-19 disease different researchers had implemented different artificial intelligence and machine learning algorithms in the past. Some of them include; The XGBoost (XGB) algorithm which was used to create a machine learning-based model for predicting survival in COVID-19 infection patients. One of the study's key findings was the model's ability to predict mortality risk with 0.95 precision and 0.90 prediction accuracy. The models give physicians a tool for identifying dangerous conditions, which helps to reduce the death rate (Yan *et al.*, 2020). Also, Sun *et al.* (2020), developed a model for predicting which COVID-19 patients will progress into critical cases. A support vector machine algorithm was used to develop the model using clinical and laboratory characteristics. The proposed model was impressive and robust in predicting patients in severe conditions, with up to 0.775 accuracies. Another observation resulted in the development of a deep convolutional neural community version with a 97.2% of accuracy for binary-type COVID-19 instances from chest X-rays (Ouchicha *et al*., 2020). In addition, with an accuracy of 98.08 % and 87.02%, the proposed model correctly detected the binary and multi-class classification of COVID-19 cases from CXIs in the study (Ozturk *et al.*, 2020).

Also, a logistic regression (LR) model to predict mortality risk among patients with severe COVID-19 was used. The most important features for distinguishing mild from severe cases have been identified as age, high sensitivity, C-reactive protein level, lymphocyte count, and d-dimer level. This model results show an accuracy prediction of 83.9% (Hu *et al.,* 2020). Another model was developed by Sánchez-Montaés *et al*., (2020). The study developed a machine learning for mortality analysis in COVID-19 patients using LR-based (Logistic regression) and machine learning techniques. The suggested model determined that age and gender were the most important factors, achieving an AUC (Area under Curve) of 0.89, a sensitivity of 0.82, and a specificity of 0.81, respectively.

Alazab *et al*., (2020), suggest an artificial intelligence-based prediction system that utilized a deep convolutional neural network (CNN) and was used to evaluate chest X-ray pictures to detect COVID-19 patients. The predicting methods could predict the numbers of COVID-19 affirmations, improvement, and death rates over the upcoming week. The average accuracy of the prediction models was 94.80 and 88.43% in two different countries. In another work conducted by Altan et al. (2020) a hybrid model that recognizes COVID-19 diseases from X-ray images using a 2D curvelet transform was created. A chaotic salp swarm algorithm and deep learning classifiers were used. The 2D curvelet transformation is applied to images obtained from X-ray radiographs of the patient's chest. The results show a correct identification of COVID-19 patients (Accuracy = 99.69%, Sensitivity = 99.44%, and Specificity = 99.81%). In another work performed by Ahmed *et.al*. (2020), COVID-19 classification was based on incomplete heterogeneous data and a KNN variant algorithm. A comparison of the variants such as Modified KNN (MKNN), KNN for imperfect data (KNNimp), and cost-sensitive KNN was provided (csKNN). The variant achieved the most streamlined performance. The result showed an accuracy of 93%.

COV-CAD, a computer-aided diagnosis (CAD) system for diagnosing COVID-19 disease in CT and X-ray datasets, was created by Ashkan *et al.* (2021). This COV-CAD system is made up of a feature extractor, a classification method, and a content-based image retrieval (CBIR) system. The feature extractor was created using a modified AlexNet CNN, and the percentages for CT and X-ray datasets are 93.20 and 99.38 %, respectively. Adi *et al.* (2021) developed a reliable convolutional neural network (CNN) model for the classification of COVID-19 based on chest X-ray views. A transfer learning-based CNN model was developed by pre-trained architectures, and each image was carefully selected to avoid bias, consisting of 368 COVID-19 pneumonia cases and 850 other pneumonia cases. In addition to assisting radiologists, the results showed that reliable COVID-19 pneumonia diagnosis from CXIs based on the CNN model opens the door to accelerating triage, saving critical time, and prioritizing resources. For the identification and classification of COVID-19, a three-stage ensemble boosted convolutional neural network for classification and analysis of COVID-19 chest x-ray images was developed by Kalaivani, and Seetharaman (2022)

Comparison of COVID-19 Prediction Techniques

A number of techniques have been proposed over the years for the predictions of covid-19. The comparison of covid-19 techniques in table 1 shows their limitations.

Table 1: Comparison of Covid-19 Techniques

Author	Strategy Used	Limitations	Datasets Used	% Accuracy
Yan *et al.,* (2020)	XGBoost (XGB)	Small dataset used	X-ray	90
Alazab et *al.,* (2020)	AI-based on Deep learning CNN	Not Available	Chest X-ray	88-94
Ahmed Hamad *et al* .,(2020)	MKNN), KNNimp, csKNN, KNN variant	Small sample size	Symptoms	93
Adi et al., 2021	CNN	Not Available	chest X-ray images	94.96
Mei-Ling & Yu-Chieh, 2022	CNN	Not Available	NIH Chest X-rays, & CT	98.33% 96.78

MATERIALS AND METHODS

This section discusses the proposed model for solving the identified problem, which is developing a machine-learning data model to predict COVID-19 in potential patients. It consists of a sequence of methods that started with the collection of data containing the parameters alongside the target variables of the Covid-19 data, the second stage implements the pre-processing procedure by removing the noises and gaps. Feature extraction was done using a combination of the univariate method (select Kbest), correlation coefficient, and extra tree classifier. The extracted features reduced the data and the number of features by eliminating insignificant and redundant features which improve the prediction accuracy. Three separate algorithms were comparatively used as the classifiers which were LR, SVM, and LR+SVM. The performance of these Algorithms was evaluated using performance metrics such as accuracy, recall, precision, and confusion matrix. Fig. 1 depicts the various stages of model development using a block diagram.

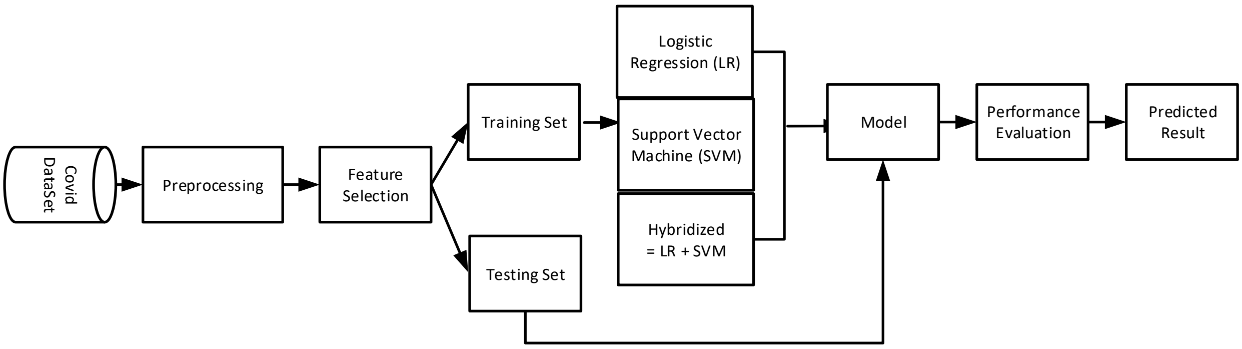


Figure 1: The Block Diagram of the Proposed Prediction model of COVID-19.

Data Acquisition

Table 2: Dataset of Symptoms.

S/N			SYMPTOMS			
1.	Running Nose	Headache	Dry Cough	Chronic Lung Disease	Breathing Problem	Sanitization from Market
2.	Asthma	Attended large gathering	Gastrointestinal	Abroad Travel	Heart Disease	Contact with COVID Patient
3.	Fatigue	Wearing Masks	Family working public exposed places	Visited Public Exposed Places	Fever	Age
4.	Hypertension	Diabetes	Sore throat	Gender		

The data used for this research work were collected from a hospital repository laboratory. The dataset which consists of 9519 characters collected from 500 individuals who were tested in the laboratories was stored in a CSV folder (table 2). It contains 23 features which include symptoms and other physical features. This allows the model to network and train various possible variations of symptoms and becomes adaptive in nature. All these entries served as a training dataset which was the input data that was fed into the model for the prediction system. Fig. 3 shows the visual representation of the dataset of features.

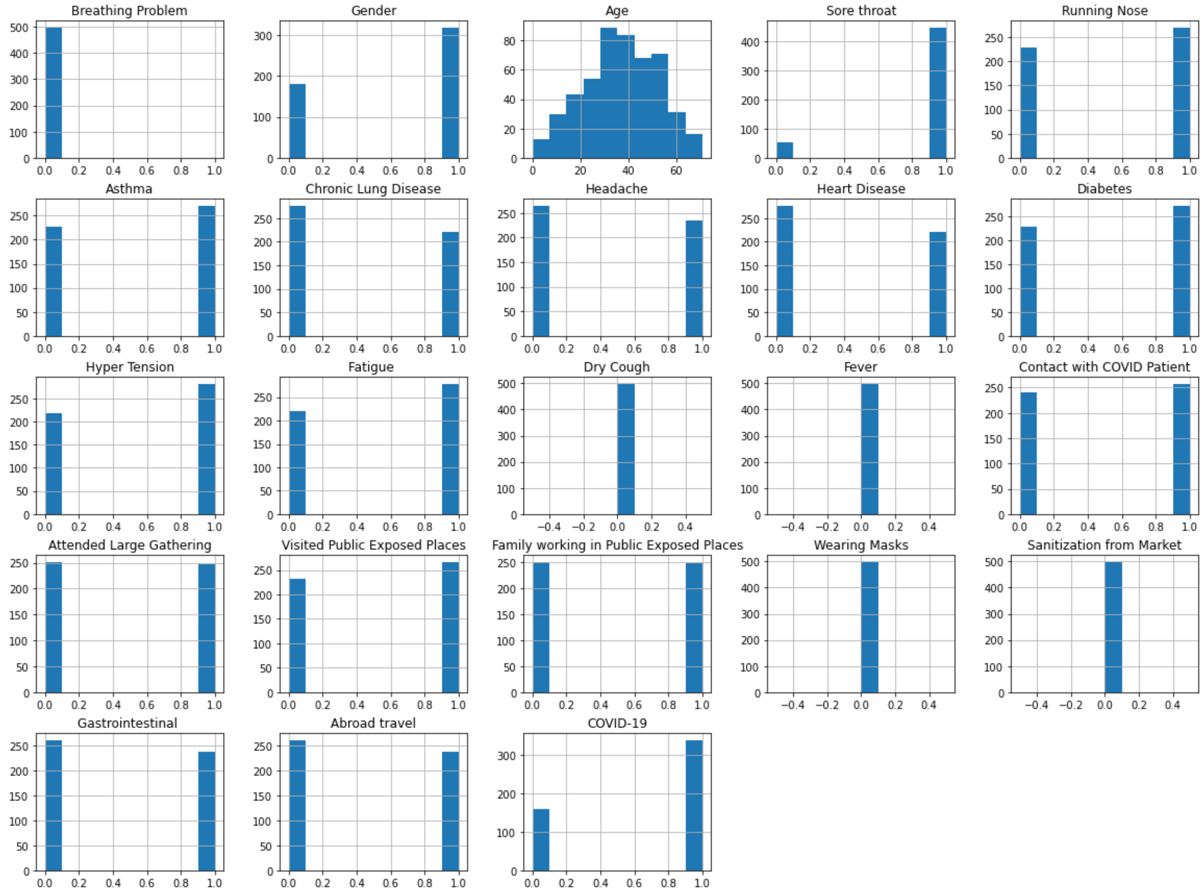


Figure 2: Visual representation of Dataset

Data Preprocessing

The data acquired were fed as input to the preprocessing stage. The raw data collected were subjected to a number of preliminary processing stages to make it usable in the descriptive stages. Preprocessing was used to remove noise and gaps from the data.

The following steps were taken in the Preprocessing stage:

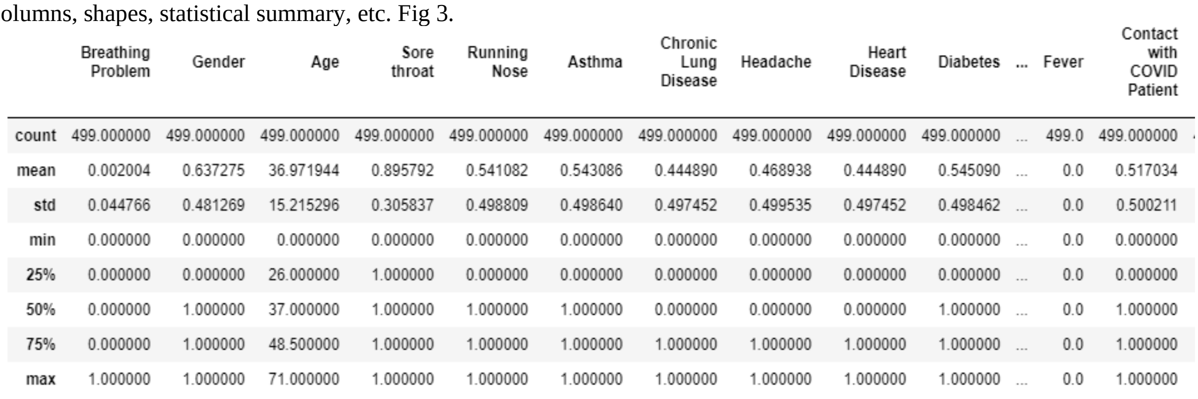
Step 1: Import the python standard libraries that are used to perform some specific jobs such as removing gaps, noise, and outliers.

Step 2: Check the dimensions of the dataset such as rows, columns, shapes, statistical summary, etc. Fig 3.

Step 3: Data cleaning for the removal of gaps and outliers, checking for duplication, handling missing data and noisy data

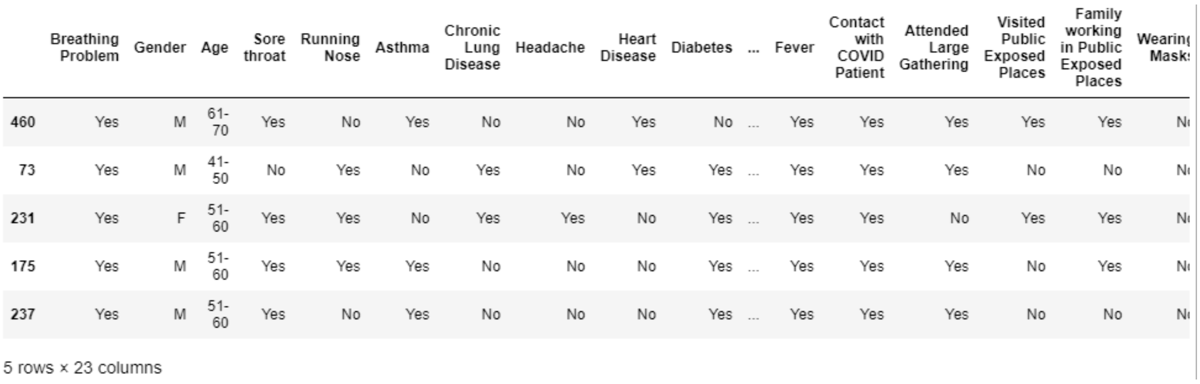
Step 4: Scaling and normalization- comprises attributes with varying scales and standardizes a dataset's independent variables into a specific range. Fig. 4 deals with categorical variables (alphabets) and their conversion to numerical variables (numbers) that machine learning can understand.

Step 5: Splitting the dataset into training and testing either ratio 80:20. It varies according to the shape and size of the dataset in question.



	Breathing Problem	Gender	Age	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fever	Contact with COVID Patient
count	499.000000	499.000000	499.000000	499.000000	499.000000	499.000000	499.000000	499.000000	499.000000	499.000000	...	499.0	499.000000
mean	0.002004	0.637275	36.971944	0.895792	0.541082	0.543086	0.444890	0.468938	0.444890	0.545090	...	0.0	0.517034
std	0.044766	0.481269	15.215296	0.305837	0.498809	0.498640	0.497452	0.499535	0.497452	0.498462	...	0.0	0.500211
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000
25%	0.000000	0.000000	26.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000
50%	0.000000	1.000000	37.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	1.000000	...	0.0	1.000000
75%	0.000000	1.000000	48.500000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	0.0	1.000000
max	1.000000	1.000000	71.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	0.0	1.000000

Figure 3: The shape and dimension of dataset.



	Breathing Problem	Gender	Age	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fever	Contact with COVID Patient	Attended Large Gathering	Visited Public Exposed Places	Family working in Public Exposed Places	Wearing Mask
460	Yes	M	61-70	Yes	No	Yes	No	No	Yes	No	...	Yes	Yes	Yes	Yes	Yes	N
73	Yes	M	41-50	No	Yes	No	Yes	No	Yes	Yes	...	Yes	Yes	Yes	No	No	N
231	Yes	F	51-60	Yes	Yes	No	Yes	Yes	No	Yes	...	Yes	Yes	No	Yes	Yes	N
175	Yes	M	51-60	Yes	Yes	Yes	No	No	No	Yes	...	Yes	Yes	Yes	No	Yes	N
237	Yes	M	51-60	Yes	No	Yes	No	No	No	Yes	...	Yes	Yes	Yes	No	No	N

5 rows × 23 columns

Figure 4: Dataset with categorical data.

Feature Selection

By extracting and refining features from raw data, feature extraction reduces the number of features or input variables in a dataset. The greater the number of features, the more difficult it is to visualize the training dataset and build a predictive model. For this research, a combination of the univariate method (select Kbest), correlation coefficient, and extra tree classifier was utilized. The steps taken were:

Step1: Import the necessary python and scikit libraries that would be used for feature extraction.

Step 2: Using the selectKbest class that can be used with a suite of different statistical tests to select a specific number of features. The chi-squared statistical test was used to select the best features.

Step 3: The correlation coefficient states how the features are related to each other or the target variable. Whether a positive correlation or negative. In this research, a heatmap was used to identify which features are most related to the target variable Figure 3.

Step 4: The last feature selection method is an inbuilt tree-based classifier that was used for extracting the best features for the dataset. The extra tree classifiers method will help to give the importance of each independent feature with the dependent feature. Each feature will give a score. The higher the score, the more relevant the feature is to the output variable depicted in Figure 4.

Step 5: Select common features from the three feature selection methods for modeling.

Classification

The next step is the development of the prediction model which employs three algorithms namely logistic regression (LR), support vector machines (SVM), and a hybrid of logistic regression and support vector machines (LR+SVM). These algorithms were chosen primarily because of their individual ability to decrease training time and increase the accuracy of prediction. The three algorithms were used sequentially herein, output from the feature extraction stages was fed into LR, SVM, and LR+SVM machine classifiers, and the output of the three algorithms was then compared using accuracy as a metric and the algorithm with the highest accuracy was used to develop the prediction system.

Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome such as yes or no based on the observations in the dataset. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. It is used when the data is linearly separable and the outcome is binary in nature (yes and no). That means logistic regression is usually used for binary classifications to predict whether the patient is infected(yes) or not(no). It can be derived from the following sigmoid function:

$$\mathrm{P} = \frac{1}{1+e} - a + bx \tag{1}$$

where P = probability, a, and b = parameters of the model.

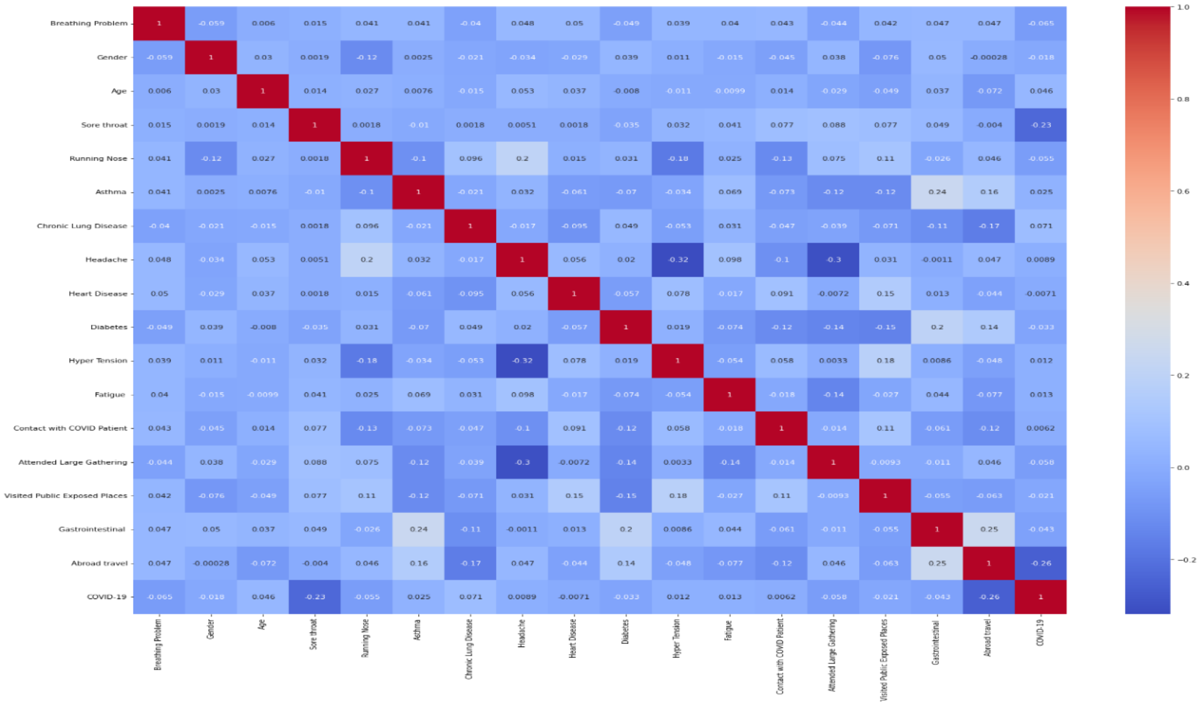


Figure 5: Heatmap of selected features

Support Vector Machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms to solve classification problems. It is a reliable classification algorithm that works well with small amounts of data. An SVM version is a point-in-area illustration of the examples which have been mapped so that the examples of the different categories are separated by as wide a gap or hyperplane as possible. Classifying medical datasets is a common procedure in machine learning algorithms like SVM. The symptoms parameters will be categorized into various groups of similar patterns, the SVM classifier will then separate the parameters into two groups of similar patterns using hyperplanes. The SVM finds the optimal hyperplane because it not only classifies the existing dataset but also predicts the class of unknown data. The optimal hyperplane is the one that has the biggest separator margin.

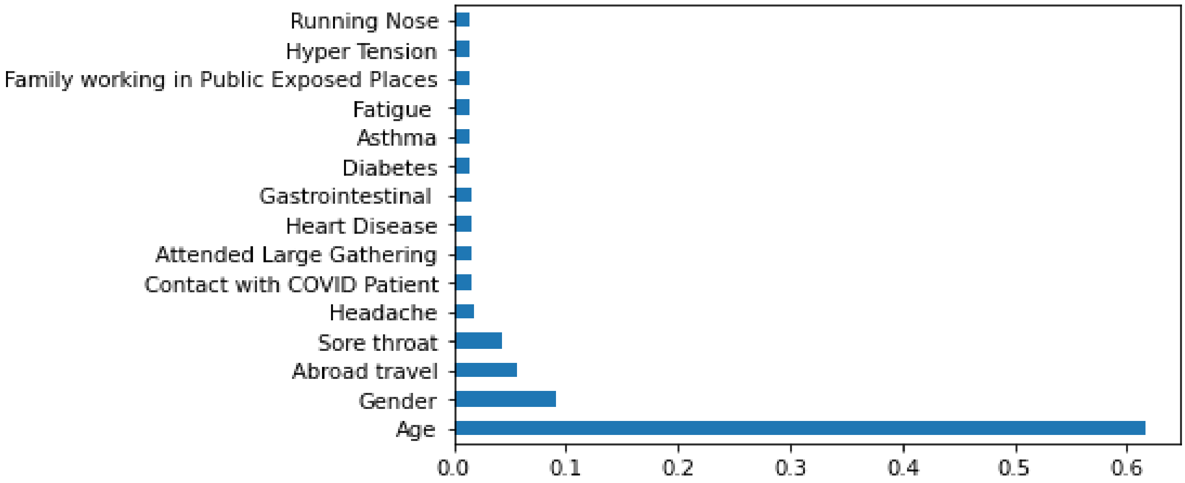
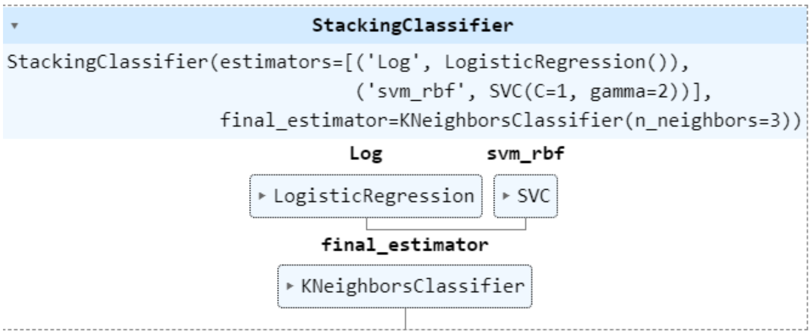


Figure 6. Graph of important features

The Hybrid LR+SVM

Logistic Regression and Support Vector Machine algorithms were hybridized using a stacking approach.

The stacking method was used to combine the techniques for logistic regression with support vector machines. In this method, a meta-learning algorithm is used to learn how to combine the predictions from two or more underlying machine-learning algorithms in the best possible way. The meta-learning algorithm for the optimization of the two models was K-Nearest Neighbor. The meta learner, which combines the prediction performance of the two algorithms, uses the individual predictions of LR and SVM as input. Fig. 5 illustrates the KNN meta-learner-based hybridization of LR and SVM for improved prediction results.



```
StackingClassifier
StackingClassifier(estimators=[('Log', LogisticRegression()),
                               ('svm_rbf', SVC(C=1, gamma=2))],
                   final_estimator=KNeighborsClassifier(n_neighbors=3))
```

Figure 7: Stacking algorithm

Evaluation

To get the parameters needed for the performance evaluation, the confusion matrix method is employed. The confusion matrix is well-suited for evaluating classification models. A confusion matrix involves the use of a two-dimensional table where the columns correspond to the predicted labels of the model while the rows correspond to the correct class labels. From the confusion matrix, we get True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Table 3 shows how these values are gotten from the table. These values form the parameters needed to calculate the performance terms of the model.

Table 3: Sample of the confusion matrix

		Predicted Values	
		Good	Bad
Actual Values	Good	True Positive (TP)	False Negative (FN)
	Bad	False Positive (FP)	True Negative (TN)

i. TP: True positives are the correctly predicted values. It is when you predict an observation belongs to a class and it does belong to that class.

ii. FP: False positives occur when you predict an observation belongs to a class when in reality it does not.

iii. FN: False negatives occur when you predict an observation does not belong to a class when in fact it does.

iv. TN: True negatives are when you predict an observation does not belong to a class and it does not belong to that class.

a) Accuracy: This gives the rate of correct predictions given by our model. It tells how often our model is right.

$$\text{Accuracy} = \frac{TP}{TP+TN+FP+FN} \qquad (2)$$

b) Precision: This metric shows how good the model is at predicting a specific category. It is used to calculate the model's ability to classify positive values correctly.

$$\text{Precision} = \frac{TP}{TP+FN} \qquad (3)$$

c) Recall: This tells how many times the model was able to detect a specific category. it is used to calculate the model's ability to predict true positive values.

$$\text{Recall} = \frac{TP}{TP+FP} \qquad (4)$$

d) F1-Score: This is the harmonic mean of recall and precision. It is useful when you need to take both precision and recall into account.

$$\text{F1-Score} = 2\times \frac{Precision \times Recall}{Precision+Recall} \qquad (5)$$

RESULTS AND DISCUSSION

The confusion matrix is a summary of prediction results on a classification problem. It involves the use of a two-dimensional table where the columns correspond to the predicted values and the rows correspond to the actual values. A confusion matrix is a good way of evaluating a good effective classification model and is used to visualize the performance of a classifier.

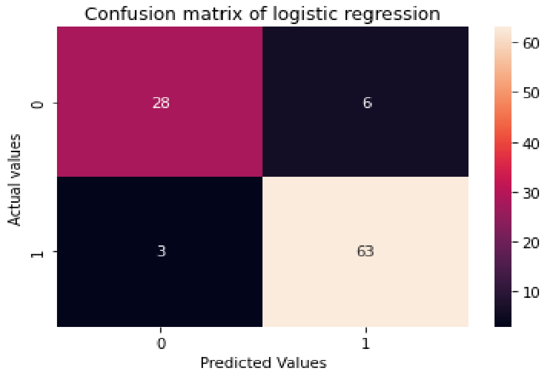


Figure 8: Confusion matrix for logistic regression

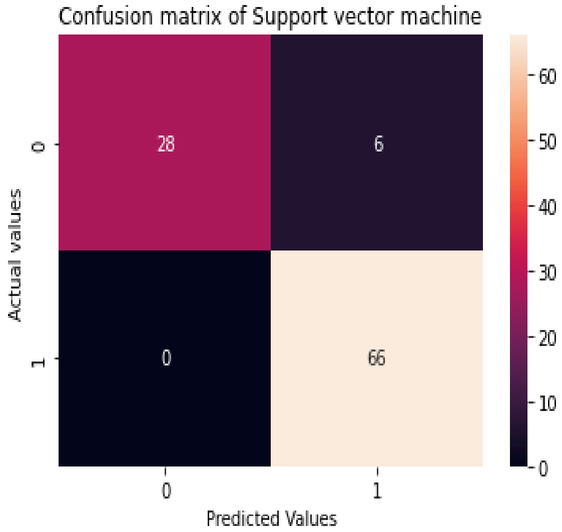


Figure 9: Confusion matrix for support vector machine

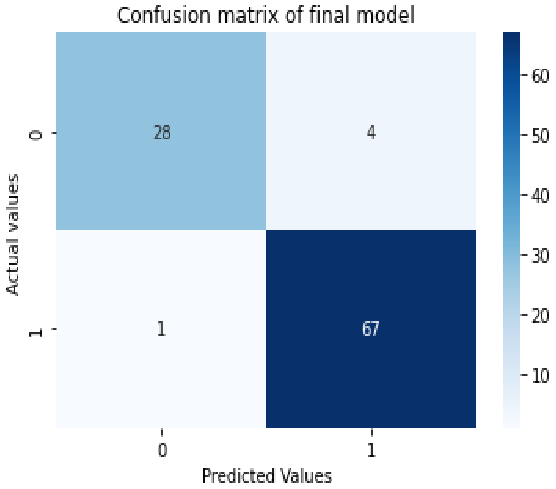


Figure 10: Confusion matrix for Hybrid Model

Figures. 8, 9, and 10 show the actual values and prediction values of both classification models from 100 samples of test data. Figure 6 shows that the model correctly predicted 28 patients with Covid-19, incorrectly predicted that 6 patients did not have Covid-19, correctly classified 66 as having no Covid-19, and correctly classified that no patients who had Covid-19 did not actually have Covid-19. Fig. 8 shows that the model correctly predicted 28 patients with Covid-19, incorrectly predicted that 4 patients did not have Covid-19, correctly classified 67 as having no Covid-19, and correctly classified that one patient who had Covid-19 did not actually have Covid-19. Table 4 shows four possible outcomes using the Confusion matrix and table 5 depicts the prediction performance of machine learning models

Table 4: Four Possible Outcomes using the Confusion Matrix

Possible Outcomes	Logistic Regression	Support Vector Machine	Hybrid Model
TP (True positive)	28	28	28
FP (False positive)	3	0	1
TN (True negative)	63	66	67
FN (False negative)	6	6	4

Hybrid = LR + SVM

Table 5: Prediction Performance of Machine Learning Models

Machine Learning	Accuracy	Precision	Recall	F1_Score
LR	91.0	94.0	92.6.	93.3
SVM	93.0	94.2	95.6	94.9
Hybrid	95.0	94.4	98.5	96.4

LR=Logistics Regression, and SVM=Support Vector Machine

Table 6. Performance Evaluation of Developed System with Related Works

S/N	Authors	Methodology	Dataset used	Performance Accuracy(%)
1	Yan *et al* (2020)	XGBoost (XGB) model	Clinical dataset	90
2	Hu *et a*l (2020)	Logistic Regression	Demographic and clinical dataset	83.9
3	Ahmed Hamad *et al* (2020)	KNN variant	Symptoms	93
4	Adi et al, 2021	CNN	Chest X rays	94.9
5	Developed Model (2021)	LR SVM Hybridized	Symptoms	91 93 95

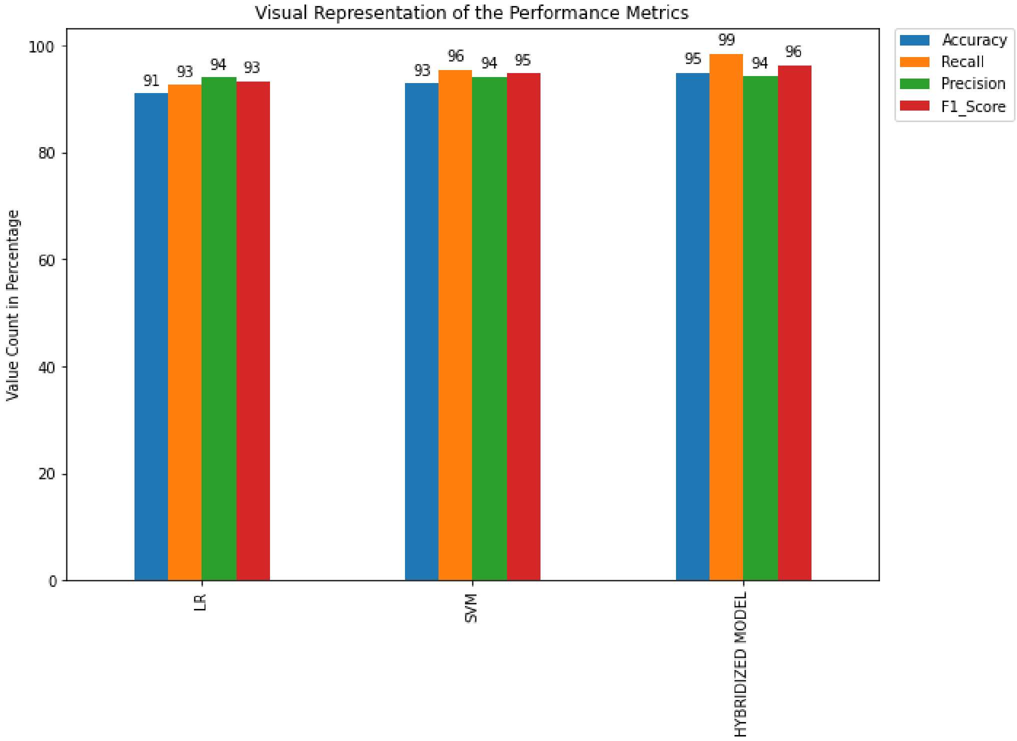


Figure 10: Visual representation of the Performance metrics

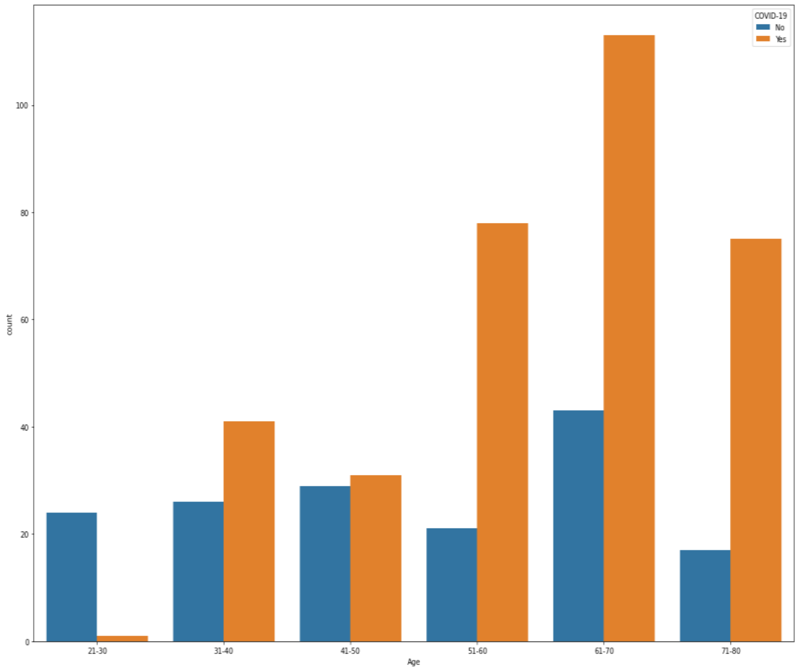


Figure 11: Gender count of Covid-19 patients

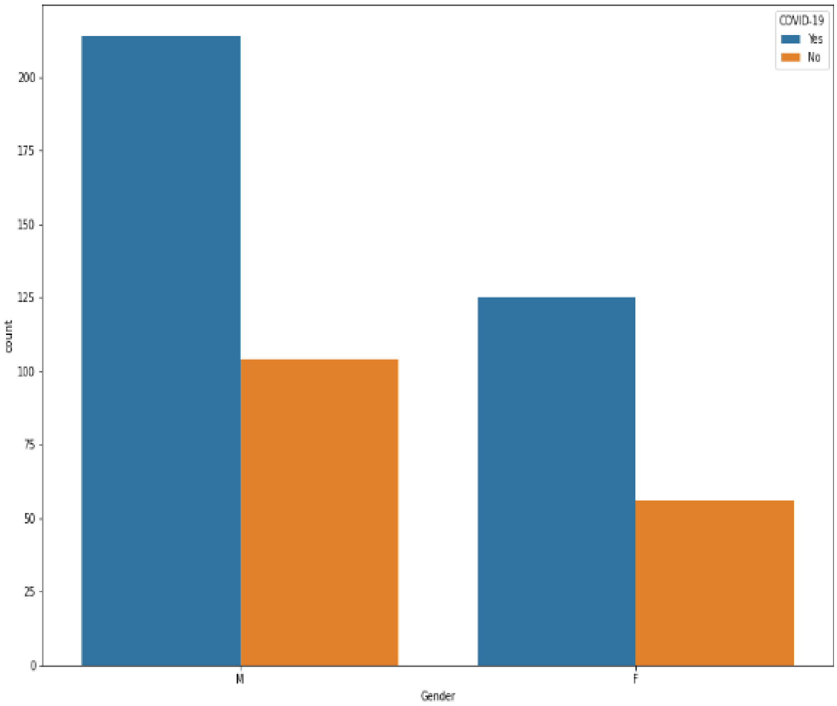


Figure 12: Age-frequency of the Covid -19 patient

According to table 6, the algorithms have an accuracy of 91% for logistic regression, 94%, for support vector machines, and Hybridized as 95%. Therefore, the Hybridized model gives an astounding result by having a 95 % accuracy, 94.4 % precision, 98.5 recall, and 96.4% F-score. Fig. 10 depicts the Visual representation of the performance metrics. Fig. 11 explains the graphical representation of the gender count of Covid-19 in the data while Fig. 12 shows the age frequency of Covid-19.

CONCLUSION

In this research work, a system was developed to predict Covid-19 using classification methods. The developed system used five main steps namely data acquisition, data preprocessing, feature selection, and classification. Data were acquired at a medical test center. Data preprocessing was applied to the Dataset to remove any noise, gaps, and outliers without losing relevant features. Feature extraction was then done on the datasets important features were selected for the prediction and irrelevant features dropped. It helps simplifies the data while retaining the relevance of the features in the data set. For the classification, logistic regression, support vector machines, and hybridized LR and SVM algorithms were used. The Hybridised model achieves astounding results of 95% accuracy, 94.4 % precision, 98.5 % recall, and 96.4 % F-score. This is so because the algorithms work in tandem to handle feature variations. As a result, a hybrid algorithm was created to highlight various symptoms that effectively identify COVID-19.

The primary purpose is to provide a fast and easy method to identify COVID-19 in patients so that action can be taken based on these results. However, the developed system or the web-based application is not meant to replace actual laboratory tests such as an RC-TCP test but to give patients a basic idea so that they can take preventative measures in the meantime before actual confirmation. It is paramount for us to make clear that under no circumstances are we trying to replace actual diagnostic tests with a classifier instead we argue that a machine learning-powered assessment system would help optimize the use of the limited testing kits and centers, especially in developing countries.

REFERENCES

Adi Alhudhaif a, Kemal Polat b,(2021) Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images, Expert Systems With Applications 180 (2021) 115141

Adigun J O, O D Fenwa, E O Omidiora, O Oladipo, SO Olabiyisi, M. M Rufai. (2015): "Development of a Genetic based Neural Network System for Online Character. Recognition", International Journal of Applied Information Systems (IJAIS) – ISSN: 22490868 Foundation of Computer Science FCS, New York, USA,Volume 9 – No.3

Adigun Oyeranmi, Babatunde Ronke, Rufai Mohammed and Aigbokhan Edwin. (2020): "Detection of Fracture Bones in X-ray Images Categorization",35(4): 1-11, 2020; Article no. JAMCS.57620

Afreen Khan and Swaleha Zubair. (2018): "Machine Learning Tools and Toolkits in the Exploration of Big Data", international journal of computer sciences and engineering, 6(12):570-575 DOI:10.26438.

Aha D.W., Kibler D and Albert M (1991):" Instance-based learning algorithms", Mach Learn,6(1):37–66.

Ahmed Hamed, Ahmed Sobhy and Hamed Nassar (2020): "Accurate Classification of COVID19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm".

Amit Y and Geman D. (1997): "Shape quantization and recognition with randomized trees", Neural Comput.,9(7):1545–88.

Anshuman Elhence, Manas Vaishnav and Shalimar. (2020): "Coronavirus Disease-2019 (COVID-19)".

Ashkan Shakarami, Mohammad Bagher Menhaj, Hadis Tarrah (2021) Diagnosing COVID-19 disease using an efficient CAD system Optik – International Journal for Light and Electron Optics 241 (2021) 167199 pp 1-12 Corresponding author. journal homepage: www.elsevier.com/locate/ij

Ashraf E., Abdallah A. and El-Sayed Atlam. (2021): "The COVID-19 pandemic: prediction study based on machine learning models".

Bracis, C.; Burns, E.; Moore, M.; Swan, D.; Reeves, D.B.; Schiffer, J.T.; Dimitrov, D. Widespread testing, case isolation, and contact tracing may allow safe school reopening with continued moderate physical distancing: A modeling analysis of King County, WA data. Infect. Dis. Model. 2021, 6, 24–35.

Cao L. (2017): "Data science: a comprehensive overview", ACM Comput Surv (CSUR),50(3):43.

Dianbo L, Leonardo C, Canelle P et al. (2020) A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models.

Elflein, J. Coronavirus (COVID-19) Disease Pandemic-Statistics & Facts|Statista. 2021. Available online: https://www.statista.com/topics/5994/the-coronavirus-disease-covid-19-outbreak/ (accessed on 30 April 2021).

Ethem Alpaydın. (2004): "Introduction to Machine Learning Publisher", MIT Press.

Fernanda Sumika, Natália Satchiko, Ben Dêivide, and de Oliveira Batista. (2020): "On the analysis of mortality risk factors for hospitalized COVID-19 patients".

Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam & Gyu Sang Choi. 2020. COVID-19 future forecasting using supervised machine learning models. IEEE Access 8: 101489- 101499.

Gao K. (2020): "Julia language in machine learning: algorithms, applications, and open issues", Comput Sci Rev 37:100254.

John G., H and Langley P. (1995): "Estimating continuous distributions in bayesian classifiers", Proceedings of the Eleventh Conference on Uncertainty in artificial intelligence,

Morgan Kaufmann Publishers Inc.; 338–345.

Harmon SA, Sanford TH, Sheng X, Turkbey EB, Roth H, Ziyue X, Yang D, Myronenko A, Anderson V, Amalou A, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. Nat Commun. 2020;11(1):1–7.

Han J, Pei J and Kamber M. (2011): "Data mining: concepts and techniques", Amsterdam, Elsevier.

Han J, Pei J and Yin Y. (2000): "Mining frequent patterns without candidate generation", ACM Sigmod Record;29: 1–12.

Hassanien E., Khaled S., Rana S., and Salloumaboul A. (2020): "Fear from COVID-19 and technology adoption: the impact of Google Meet during Coronavirus pandemic", DOI:10.1080/10494820.2020.1830121.

Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl. 1998;13(4):18–28. Hu C, Liu Z and Jiang Y. (2020): "Early prediction of mortality risk among patients with severe COVID-19, using machine learning", International Journal of Epidemiology, vol. 49, no. 6, pp. 1918–1929.

Jackins, V., Vimal, S., Kaliappan, M. & Lee, M.Y. 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of

Supercomputing 77: 5198-5219. https://doi.org/10.1007/ s11227-020-03481.

Jason Brownlee," Machine Learning Tools", December 28, 2015.

Das, Rabi Narayan Behera. "A Survey on Machine learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering. vol. 5,

Kalaivani, S and Seetharaman, K (2022) A three-stage ensemble boosted convolutional neural network for classification and analysis of COVID-19 chest x-ray images, International Journal of Cognitive Computing in Engineering 3 (2022) 35–45

Kaelbling L.P., Littman M.L. and Moore A. W. (1996): "Reinforcement learning: a survey", 4:237–85.

Kamble S.S., Gunasekaran A and Gawankar S.A. (2018): "Sustainable industry 4.0 framework: a systematic literature review identifying the current trends and future perspectives", Process Saf Environ Protect,117:408–25.

Keerthi S.S., Shevade S.K., Bhattacharyya C, Radha Krishna M.K. (2001): "Improvements to Platt's SMO algorithm for SVM classifier design". Neural Comput,13(3):637–49.

Khadse V., Mahalle P.N. and Biraris S.V(2018): "An empirical comparison of supervised machine learning algorithms for internet of things data", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, 1–6.

Lalmuanawma S, Hussain J and Chhakchhuak L (2020): "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: a review", Chaos Sol Fract,110059.

LeCessie S, Van Houwelingen J.C. (1992): "Ridge estimators in logistic regression". J R Stat Soc Ser C (Appl Stat),41(1):191–201.

Liu H and Motoda H. (1998): "Feature extraction, construction and selection: A data mining perspective", vol. 453. Springer Science & Business Media.

Machine Learning Model Development and Model Operations: Principles and Practices,.2021 https://www.kdnuggets.com/2021/10/machine-learning-model-development-operations-principles-practice.html.

Martinez-Velazquez, R.; Tobón, V.D.P.; Sanchez, A.; El Saddik, A.; Petriu, E. A Machine Learning Approach as an Aid for Early COVID-19 Detection. Sensors 2021, 21, 4202.

Morens DM, Daszak P, Taubenberger JK. Escaping Pandora's Box—another novel coronavirusexternal icon. N Engl J Med. 2020. https://doi.org/10.1056/NEJMp2002106.

Mei-Ling Huang, Yu-Chieh Liao (2022) A lightweight CNN-based network on COVID-19 detection using X-ray and CT images, Computers in Biology and Medicine 146 (2022) 105604, pp 1-13

Miao J. and L. Niu, "A survey on feature selection," Procedia Computer Science, vol. 91, pp. 919–926, 2016.

Alazab,Albara Awajan, Mesleh Ajith and Abrahama, (2020): "COVID-19 Prediction and Detection Using Deep Learning".

Muhammad LJ, Usman SS (2020) Power of artificial intelligence to diagnose and prevent further COVID-19 outbreak: a short communication (2020); arXiv 2004.12463.

Nemati, M.; Ansary, J.; Nemati, N. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. Patterns 100074.

Ouchicha, C., Ammor, O., & Meknassi, M. (2020). CVDNet: A novel deep learningarchitecture for detection of coronavirus (Covid-19) from chest x-ray images. Chaos,

Solitons & Fractals, 140, 110245. https://doi.org/10.1016/j.chaos.2020.110245

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X- ray images. Computers in Biology and Medicine, 121, 103792. https://doi.org/10.1016/j.compbiomed.2020.103792

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R and Dubourg V. (2011): "Scikit-learn: machine learning in python", J Mach Learn Res.; 12:2825–30.

Roosa, K., Lee, Y., Luo, R., Kirpich,A., Rothenberg,R.,Hyman, J.M., Yan, P. & Chowell, G. 2020. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th. Infectious Disease Modelling 5: 256-263

Sarker I. H., Kayes ASM., Badsha S., Alqahtani H. and Watters P. (2020): "A. Cybersecurity data science: an overview from machine learning perspective", J Big Data.,7(1):1–29.

Sarker I. H., and Kayes ASM. (2020): "user behavioral rule-based machine learning method for context-aware intelligent services", J Netw Comput Appl.; page 102762.

Sarker I.H. (2019) "A machine learning-based robust prediction model for real-life mobile phone data", Internet Things,5:180–93.

Sánchez-Montañés, Rodríguez-Belenguer, and Alakhdar-Mohmara, Y. (2020) "Machine learning for mortality analysis in patients with COVID-19," International Journal of Environmental Research and Public Health, vol. 17, no. 22, pp. 8386–20.

Sanchez-Maro, A. Alonso-Betanzos, and M. Tombilla-Sanrom an,"Filter methods for feature selection a comparative study," in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2007.

Seçkin Karasu, Bülent Ecevit, Üniversitesi Aytac, Altan.(2020): “Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique”, DOI:10.1016/j.chaos.2020.110071.

Sharma, A., Tiwari, S., Deb, M.K. & Marty, J.L. 2020. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): A global pandemic and treatment strategies. International Journal of Antimicrobial Agents 56(2): 106054. https://doi. org/10.1016/j.ijantimicag.2020.106054.

Shirzadi A. (2018): “Novel GIS-based machine learning algorithms for shallow landslide susceptibility mapping”, 18(11):3777

Simionattoa A., Varonaa H. and Panjehc L. (2020): “Fractal signatures of the COVID-19 spread”, Volume 140.

Srivastava, M. N. Joshi, and M. Gaur, “A review paper on feature selection methodologies and their applications,” IJCSNS, vol. 14.

Sun L, Song F, Shi. N. (2020), “Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19,” Journal of Clinical Virology, vol. 128, p. 104431

Topol E., “High-performance medicine: the convergence of human and artificial intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019.

Teresa Lambe, Neeltje, and Van Doremalen (2020): “CoV-19 vaccine prevents SARS-CoV-2 pneumonia in rhesus macaques”.

Venkatesh and J. Anuradha, “A review of feature selection and its methods, “Cybernetics and Information Technologies, vol. 19, no. 1, pp.3–26, 2019

Wu, Yi-Chi, Ching-Sung, Chan, and Yu-Jiun (2020): “The outbreak of COVID-19: An overview”, Journal of the Chinese Medical Association, Volume 83, p 217-220.

Yang XiaoLi, Yan, Hai-Tao Zhang and Jorge Goncalves. (2020): “An interpretable mortality prediction model for COVID-19 patients”, Nature Machine Intelligence Volume 2, pages 283–288.

Yildirim, “Filter-based feature selection methods for prediction of risks in hepatitis disease,” International Journal of Machine Learning and Computing, vol. 5, no. 4, p. 258.

Mohanasundaram,A & Aruna,S.K.(2022) “Improved Henon Chaotic Map-based Progressive Block-based visual cryptography strategy for securing sensitive data in a cloud EHR system” International Journal of Intelligent Networks Volume 3, Pages 109-112