THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 3. Issue 1 (2023)
https://doi.org/10.62019/abbdm.v3i1.45.

ASIAN BULLETIN OF BIG DATA MANAGEMENT
ISSN (Print): 2959-0795
http://abbdm.com/
ISSN (online): 2959-0809

Prediction of COVID-19 using machine learning techniques

Muhammad Afzal Raja*, Inam Ullah, Muhammad Babar, Tuba Aziz

Chronicle

Article history
Received: October 28, 2023
Received in the revised format: Nov,5 2023
Accepted: Nov 07, 2023
Available online: Nov 15, 2023

Muhammad Afzal Raja is currently affiliated with Department of Computer Science, Mohi-Ud-Din Islamic University, Nerian Sharif, Azad Jammu Kashmir, Pakistan
Email: afzalghaias@gmail.com

Inam Ullah (PhD), is currently affiliated with Department of Computer Science, Mir Chakar Khan Rind University Sibi, Balochistan, Pakistan.
Email: sahinam@gmail.com

Muhammad Babar & Tuba Aziz, are currently affiliated with Department of Computer Science, Mohi-Ud-Din Islamic University, Nerian Sharif, Azad Jammu Kashmir, Pakistan.
Email: babarsardar21@gmail.com
Email: Tuba.akbar513@gmail.com

***Corresponding Author**

Abstract

In December 2019, the world faced an unprecedented and formidable challenge in the form of COVID-19. Since its first reported case in December 2019 and subsequent classification as a pandemic by the World Health Organization (WHO), it has imposed an enormous impact, taking lives, disrupting diverse economic sectors, and introducing numerous challenges. Predicting and controlling COVID-19 precisely remains a pivotal concern for the future. To enhance the precision of COVID-19 disease prediction and alleviate the burden on healthcare systems, we explore the application of diverse machine learning classifiers. Leveraging datasets comprising confirmed cases, recoveries, and fatalities, our study seeks to improve the predictive accuracy, ensuring efficient and precise evaluation and triage of patients. This, in turn, the load on overburdened emergency departments is reduced, and mitigates the pressures faced by healthcare professionals. The pivotal role of artificial intelligence (AI) and machine learning (ML) in this context cannot be overstated. Our research endeavors to refine the accuracy and quality of results by employing advanced machine learning techniques such as Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost). Our dataset is sourced from Kaggle, a well-established platform for data science and machine learning. Our comprehensive analysis compares the outcomes of six distinct models. Notably, the Gradient Boosting classifier surpasses other techniques, achieving an impressive accuracy rate of 90% for confirmed cases, 90% for recoveries, and 92% for fatalities. This represents a significant improvement over the baseline paper, which achieved accuracy rates of 83% for confirmed cases, 72% for recoveries, and 81% for fatalities using the same dataset. Further our research enhances COVID-19 prediction, aiding healthcare professionals in addressing the other global epidemic for future.

Keywords: COVID-19, Machine learning, Data Mining, Data Science, Artificial Intelligence.

INTRODUCTION

COVID-19, a novel strain of the Coronavirus, had spread considerably by the end of 2019 (Brunese est al., 2020). A new strain (COVID-19) acronym for Coronavirus Disease 19, was detected in humans that had never been previously detected (C. Huang et al., 2020). In December 2019, people in Wuhan City, China, became the first to be infected with the virus (Khanday et al., 2020). After one of the cases was diagnosed as a new virus called COVID-19, China released the genetic sequence of the virus to the public on 12 January 2020 (She et al., 2020). On 30 January 2020, the World Health Organization (WHO) declared the epidemic an emergency in public health of worldwide concern, then on 11 March 2020, it declared the outbreak a pandemic. (Kumari et al., 2021). Coronaviruses are a kind of respiratory virus that has the potential to cause a variety of disorders, MERS (Middle East respiratory syndrome) and SARS (severe acute respiratory syndrome) are two examples of respiratory syndromes. (Holmes, 2003). They acquire their name from the crown-shaped points that sprout on their surface (Van Der Hoek et al., 2004). These viruses are common in many animal species such as camels and bats, but they can also develop and infect humans, potentially spreading to the population (Brunese et al., 2020). The COVID-19 outbreaks not only harm people's lives, but they also have a profound effect on the country's economy. As of April 28, 2020, there was no vaccination to prevent this virus from infecting more than 3 million individuals. Precautionary measures were issued by the WHO. Using information technologies, deep learning and machine learning (ML) can help combat this epidemic. (Painuli et al., 2021). On August 22, there was evidence of human-to-human transmission, and on August 30, it was granted a high-risk designation (She et al., 2020). The fact that COVID-19 was discovered and the subsequent pandemic was an unanticipated occurrence with few to no or unknown options for recognizing, handling, or regulating sick people explains why it has had such an impact. However, during the past two years, we have gotten used to it and are now beginning to regain our normal lives as they were prior to the epidemic. To be able to adapt to any emerging diseases that may arise in the future, we must continue to possess the capabilities to evaluate ailments like COVID-19 adequately. According to (Samuel, 1959), machine learning is the branch of science that enables computers to learn without being supervised.

As a result, we can define ML as the field of computer science in which computers that can program themselves can be constructed. Hospitalized individuals with the coronavirus syndrome (COVID-19) are constantly in danger of death. ML algorithms might be used to predict death in COVID-19 hospitalised patients. As a result, our research will evaluate several ML algorithms for forecasting using information from individuals at the time of admission for COVID-19 morbidity, with the purpose of selecting the much more appropriate algorithm for use in formulating decisions. Artificial intelligence-based electronic gadgets can be crucial in controlling the spread of the virus as it spreads from person to person. The importance of electronic health data has expanded along with the function of healthcare epidemiologists (Bates et al., 2014). A significant opportunity for medical research and useful applications to advance healthcare is presented by the growing accessibility of electronic health data. This information can be used to improve the decision-making capabilities of machine learning algorithms in terms of disease prediction (Wiens & Shenoy, 2018) . Pattern recognition, a branch of artificial intelligence (AI), gave rise to machine learning, where data can be organized for user understanding.

Many applications have recently been developed utilizing machine learning across a multitude of areas, including healthcare, finance, and military hardware, space exploration, etc. Machine learning may be implemented in healthcare to provide enhanced diagnostic tools for medical image analysis. For example, a machine learning algorithm can be used in medical imaging to look for patterns that indicate the presence of a specific illness using pattern recognition. This might help doctors make more quickly and accurately diagnosis. (Wernick et al., 2010). Machine learning could potentially be used to analyse clinical trial data in order to discover previously unknown medication adverse effects. This has the potential to optimise patient care and also enhance the safety and efficacy of medical operations. ML is a rapidly growing and developing field. It optimizes the computer's performance using data-driven programming. It uses the training data or its prior experiences to learn the parameters to optimize the complex algorithms (Venkata Ramana et al., 2011). It can also make future predictions using the data. Using the statistics of the data, machine learning also contributes in the construction of a mathematical model.

LITERATURE REVIEW

Before beginning the research process, it was necessary to review previous works in the field to obtain a broad knowledge of the findings that are already available in this domain. The approaches currently employed in research on COVID-19 prediction are discussed in this literature review. The current machine learning and epidemiology models were the primary focus of this preliminary research, with the goal of identifying a potential research gap and eventually strengthening this area of research with our own work. (Muhammad et al., 2021) used DT, NB, SVM, logistic regression and ANN ML algorithms to develop models of supervised ML for the positive and negative of COVID-19 cases in Mexico. 80:20% data is used for training testing. The DT algorithm outperforms than the NB and logistic regression ML algorithms in terms of accuracy. By using a decision tree model, it is also indicated that the age factor plays a significant role from the dependent variables. People over the age of 45 are more likely to be infected with SARS-CoV-2 than those under the age of 45. They also state that patients with diabetes, obesity, hypertension, asthma, and pneumonia have a much higher risk of contracting COVID-19 than the other patients.

(Mary & Albert Antony Raj, 2021) determines which classification strategy, in relation to the data samples for COVID-19 patients diagnosed, operates with high precision. In this research the author used classification methods including NB, KNN, DT, RF and SVM with a linear kernel (linearly separable) for maximum accuracy. For classification the author used both numerical and categorical variables. Conversely, it was found that SVM is the best and has the most elevated precision pace of 85%, which is exceptionally helpful for COVID-19 clinical consideration with a little information assortment. (Lasya et al., 2022) use numerous algorithms for ML prediction models and then computed the performance and evaluation of these models. Authors compare the results of different models including Multilinear Regression, LR and XGBoost Classifier, RF Regressor and RF classifier, SVM, DT Classifier, NB Classifier and KNN+NCA and finally concluded that the performance of RF Regressor and FF Classifier is outclass.

(Arpaci et al., 2021) used six different kind of classifier including PART, J48, IBk, BN, CR and Logistic by using 14 medical features to develop predictive model. The author use COVID retrospect 114 cases from a hospital which is situated in china. Finally they concluded that the result CR - meta classifier is the best with 84% accuracy for predicting corona positive and negative cases. Since the spread of novel COVID-19 is impacted by many variables, various sorts and intensities of intervention will bring totally different outcomes. Consequently, a simple framework can't be used for complete prediction. Novel COVID-19, as a recently arising infectious disease, is hard to anticipate its pandemic pattern when using sophisticated models with only a small number of parameters. So, a BP neural network model with fewer parameters is advantageous in demonstrating strategies with comparable execution. (Zhao et al., 2021). COVID-19 is a significant concerning matter for all over the world. The latest daily basis data is collected from the website of university of johns Hopkins and applied the ARIMA model on COVID-19 information that is from January 20, 2020, to February 10, 2020, to predict the COVID-19 cases all over the globe for the next two days. They have used partial autocorrelation (PACF) and the autocorrelation work (ACF) graph to pick the optimal model characteristics. (Benvenuto et al., 2020).

The primary purpose of (Daniyal et al., 2020) is to offer a prediction approach that may be used to anticipate Pakistan's COVID-19 mortality case trend. Using a descriptive statistical analysis, the victims of COVID-19 by age and gender were depicted. In this work, three regression models—logarithmic, linear, and quadratic—were used to simulate COVID-19 mortality instances in Pakistan. On the basis of R^2, adjusted R^2, AIC, and BIC criteria, these three models were analyzed. The National Institute of Health of Pakistan provided the data used in the modelling, which covered the period from February 26, 2020, to August 5, 2020. The estimated data from all three approaches were shown against the observed data. Comparatively speaking to the other two models, quadratic regression exhibits a greater match. (Painuli et al., 2021) purposed a model for anticipating COVID-19 on the basis of symptoms, described by the CDC and the WHO. The author constructed a catalogue of symptoms into which standards were built and entered. These statistics were then utilized as raw data. The next step in organizing the data was feature extraction. They used ARIMA time series data to predict confirmed cases in various Indian states.

The data were split into training data (80% of the data) and test data (20% of the data). They choose two techniques, RF and ETC, both of which are more than 90% accurate. ETC's accuracy rating is 93.62%, which is higher than RF's. In order to improve future work, additional factors and techniques can be used with ARIMA to get more precise results. Azarafza et al. (2020) use LSTM neural network for the prediction of COVID-19 at national and provincial levels in Iran and is used for time series modelling. For confirmed cases of COVID-19, the LSTM is applied. According to the Iranian Department of Health and Medical Education, the data included in the model was collected at the state level between February and March 2020. It also compared LSTM to seasonal ARIMA, moving average, and exponential smoothing approaches, and determined that LSTM outperformed the others. Khanday et al. (2020) used ensemble and standard ML techniques to classify linguistic clinician reports in this study into four categories. In feature engineering, the terminology bag of words (BOW), frequency/inverse features are expected (TFR/IDF), and reporting duration were employed. They collected data of 212 individuals' and information is archived from the open-source data repository GitHub, along with their corona virus and other viral symptoms.

There are approximately 24 characteristics in data. Pipeline is being used to improve the accuracy of all ML algorithms. The data is divided in a 70:30 ratio, with 70% of the data used for training the model and 30% used for testing the model. It was discovered that logistic regression and multinomial Naive Bayesian classifiers produce outstanding results with 94 percent precision, 96 percent recall, 95 percent f1 score, and 96.2 percent accuracy. Other ML techniques that performed well were RF, SGB, DT and GB. Mandayam et al. (2020) employed two supervised-learning models to generate estimations using the time series dataset for COVID-19. For the purpose of conducting experimentation in supervised learning, two regression models, LR and SVM were utilized. The fundamental dataset consists of 157 attributes and 266 tuples. Depending on locations, the dataset contains all the positive instances that were reported and dated from 01/22/20 to 06/22/20. The dataset is then divided into train and test data using the following ratios: 30:70, 50:50, 60:40, and 80:20. For more accurate results, the author divided the data set into 85:15 ratio. Finally, it is concluded that the LR technique performs better. According to (C.-J. Huang et al., 2020) more than 346,000 confirmed cases and more than 14,700 deaths had occurred up until March 23, 2020, in 173 different nations and regions of the world. Over 81,000 confirmed cases and over 3,200 fatalities were reported outside of China. This study concentrated on several cities with the highest number of confirmed Chinese instances, along with a COVID-19 prediction method relying on the Convolutional Neural Network (CNN) approach. The measures of MAE and RMSE were used to compare the overall efficacies of various algorithms. The experiment results showed that the CNN model proposed in this study has the highest prediction efficacy when compared to other deep learning techniques. Pourhomayoun & Shakibi (2021) employed 307,382 labelled samples of both male patients and female patients, with a mean lifespan of 44.75, from a collection with more than 2.67 million COVID-19 patient populations from 146 countries with test verification. A total of 112 parameters, including ailments, medical notes from physicians, demographic data, and biological data, were extracted out of the original dataset. Several ML models including SVM, ANN, RF, DT, LR, and KNN to compute the COVID-19 patient fatality rate. The results show that using neural network model, the assumptions about the fatality rate were generally accurate to 89.98%.

(Sujath et al., 2020) offered a potential model to predict the spread of COVID-2019. The author applied linear regression, multilayer perceptron, and vector auto regression methods on the COVID-19 dataset. In the case of multivariate time series, the VAR model outperformed than other estimation techniques and is increasingly used in practical anticipating situations. Sun et al. (2020) made an effort to find the most suitable optimization technique for early COVID-19 recognition using clinical huge data including 912 participants from 18 clinics in Zhejiang. Five different classifiers including SVM, LR, DT, RF, and DNN were used. For the early COVID-19 quick screening, the LR algorithm is demonstrated to be the best strategy out of the five identification alternatives. Employing four ML techniques including SVM, NB, KNN, and RF. They categorized COVID-19 instances using human nucleotide DNA. Public coronavirus genotypes are gathered for the 2019 Novel Coronavirus Resource (2019nCoVR) by China's National Center for Bio information from a variety of databases, namely NCBI, NMDC, GISAID, and CNCB/NGDC. The practice of tenfold cross-validation was used. The accuracy of the KNN and RF strategies using genetic sequence traits was 92% and 93%, respectively, according to experimental observations.

(Rochmawati et al., 2020), used a DT method is to characterize the ailments in a symptomatology dataset. J48 and Hoeffding Tree were employed on Weka. The data were divided into two groups: training data (66%), and testing data (34%). The HT and technique J48 are both relatively unremarkable. Just in terms of accuracy, precision, and recall, the J48 scores in this trial were marginally superior to those of the HT. The HT is less complex and has fewer nodes than J48, according to the results of the tree view. Fayyoumi et al. (2020) collected larger dataset from common survey. Individuals who were willing to participate in the study were asked if they passed the eligibility requirements. This data served as the input for Logistic and Linear Regression, Support Vector Machine and Multi-Layer Perceptron. These algorithms were used to identify prospective COVID-19 patients based on their symptoms and indicators. When tried to compare to the other models, the MLP has demonstrated the highest accuracy (91.62%). The SVM, on the other hand, has the highest precision (91.67%).

(Kumari et al., 2021) provides a comprehensive examination of newly formed estimation methods and calculates the number of verified, recovered, and fatal COVID-19 cases in India. Auto - correlation and auto regression have been implemented to improve accuracy along with regression coefficient and multiple linear regression employed for prognosis. They anticipated number of scenarios and the actual values correspond well (0.9992 R-squared score). Based on verified, recovered, and fatal cases reported in India, the COVID-19 dataset is examined for coronavirus illness. The following methodology is employed to make accurate predictions for data from 20th of the March 2020 to 6th of the June 2020. This dataset's features primarily took confirmed, recovered, and death cases into consideration. Gupta et al. (2021) conducted analyses based on cases that occurred in India's various states in chronological order. The dataset includes COVID-19 data features that were obtained from the Ministry of Health and Family Welfare and from Kaggle. The dataset contains only 2342, COVID-19 cases from India between January 30, 2020 and May 26, 2020. Researchers are using several divisions of categorization because the dataset contains various classes. Among these classifiers RF, LR, SVM, DT, and KNN, the RF having the highest accuracy of 83.54%. The random forest is utilized for outcome analysis and prediction. The K-fold cross-validation is used to evaluate the model's consistency.

The COVID-19 dataset retrieved from John Hopkin's university-repository was used by (Gothai et al., 2021) .The data was collected in 2020 between January and December. For future prediction supervised machine learning methods including Time-series Holt's model, LR and SVR were used. In comparison to LR and SVR algorithms, this work provided a time series forecasting Holt's winter model that has superior precision in predicting future data with 87% accuracy.

METHODOLOGY

The most essential challenge in this research is to design the best predictive model that can predict COVID-19 more correctly. A tremendous amount of work is being done in this area, but due to the spread of the pandemic, there is still a need to build more accurate and efficient system. Because COVID-19 prediction plays such a vital role in our daily lives, researchers are working hard to enhance it. Various strategies have been used to predict COVID-19. One of them is "Data Mining," which produces reliable outcomes. We applied various data mining approaches to find a better prediction system.

Proposed Model

Different approaches will be used in the research in order to accomplish the research goals and objectives. The methodology is depicted in Figure

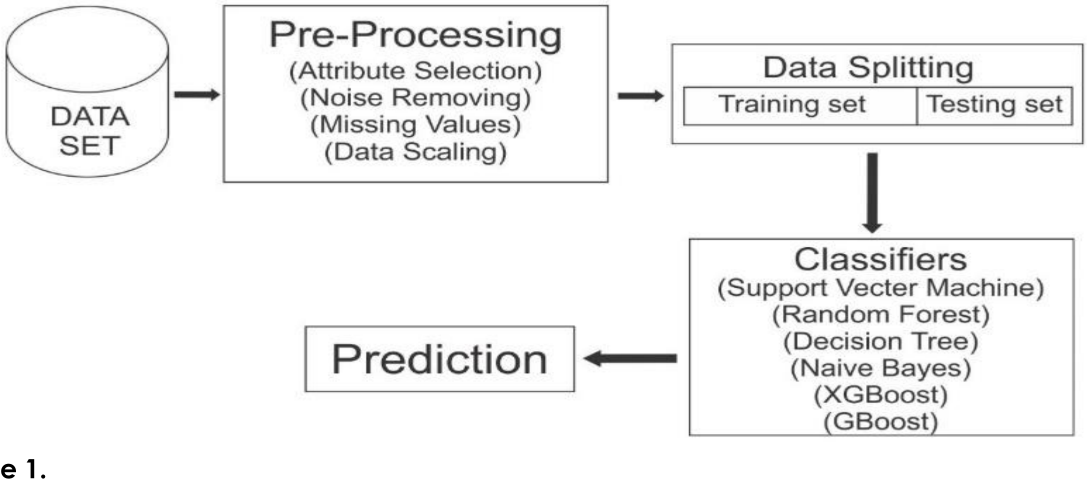

Figure 1.
Proposed Context Diagram/ Architecture

Domain Selection

Although a lot of work has previously been carried out in this area, COVID-19 prediction is one of the most concentrated research topics. A number of prediction techniques are used to forecast the COVID-19. We have selected this area for our research by considering the significance of a precise COVID-19 prediction. The primary datasets for our study project were obtained from the KAGGLE website, which is the most reliable source for real-world datasets.

Data Source Identification

The first stage involves identifying several data sources for the gathering of COVID-19 prediction related facts. The data set utilized to develop and train the COVID-19 prediction model, was obtained from open-source data made available by Kaggle and IHO. Information on the problem was gathered from various sources and saved in the system for future use. In the current stage, data was collected in raw form; we must preprocess it for the later steps. After the identification of the data sources, we move to the next stage, which is data collection.

Data Collection

The act of assembling, measuring, and evaluating an accurate data set for research purposes utilizing conventional verification methodologies is referred to as data collecting. We acquired datasets from KAGGLE for this research work because a researcher can evaluate his theory based on the data sets collected. The data set contains information about COVID-19 hospitalized patients. It comprised demographic information, admission and discharge dates, the number of deaths and cures, the location, age and gender derived from computerized records. All the attributes have been eliminated which are not necessary for our model. The data-set consists of multi-dimensional data that has been combined. It has textual data fields as well as numerical values.

Table 1.
Data set attributes

Attribute no/ feature no	Attribute name/ feature name
1	Date
2	Time
3	State / UnionTerritory
4	Confirmed Indian National
5	Confirmed Foreign National
6	Cured-COVID-cases
7	Deaths-COVID-cases
8	Confirmed-COVID-cases

For the purposes of the experiment, Integer values were used to encode text data. Table: 1 shows the attributes which were examined in the dataset corresponding to the ML model.

Implementation Programming Language

The programming language we'll be utilizing for this research is Python. When it was initially launched in 1991, Guido van Rossum gave it its initial introduction. The rising use of Data Science over the past few years has been credited with the Python programming language's surge in popularity. Python is an interpreted language for object-oriented programming with a high level of abstraction that is utilized in many different contexts, including data analysis projects, website backend code, and scientific research.

Google Colab

We make use of the Google Colab for this research. Google Research created the"Colab" product. With Colab, anyone can write and execute arbitrary Python code in the browser, making it a fantastic tool for machine learning, data analysis, and all forms of education. Google Colaboratory offers free usage of Jupyter Notebooks. Given that it is a Google cloud-based service, using it is totally free. One benefit of Colab is that there is no need for pre-installation. As above mentioned that this service is an excellent free rendition of a hosted Jupyter notebook that does not required any configuration and gives users exposure to Google's computational capabilities, particularly GPUs and TPUs. When the browser is closed, a Colab's instance can run for up to 12 hours and 90 minutes before being declared idle and being recycled. With Google Colab, you can study and construct Python machine learning models quickly. It is built on the Jupyter notebook and allows for collaboration. The notebooks can be shared and edited by the team members even when they are separated geographically. It is also possible to post the notebooks on GitHub and make them available to the broader public. Many well-known machine learning frameworks, such as PyTorch, TensorFlow, Keras, and OpenCV, are supported by Colab.

Data transformation

Data preparation and transformation are steps in the data mining process that aim to increase the effectiveness of knowledge discovery. The first step in data preprocessing is to collect only relevant dataset components that are helpful to this research work. Dataset components that are unrelated to our research, such as inconsistent and lacking in clear behaviors or trends are discarded before being stored. After that we converted textual data into numeric by using transformation functionality of PYTHON.

Classification

Classification is the process of categorizing data in order to give it a defined form and cohesive structure that allows it to be applied in the most systematic and efficient possible way. Classification is the technique of dividing the data into different groups. It is the process of aggregating statistical data into separate homogeneous groups that allows for easy interpretation. In classification first we will apply NB, DT, RF, SVM, GB and XGBoost classifier on data sets. After classification we moved to testing and evaluation.

Model Building

It is challenge in constructing a model for the COVID-19 prediction. The first and most crucial element for this is to have background knowledge about COVID as well as about prediction models. As COVID-19 begins in start of 2020, various methods and prediction models are employed. The model that will be used to predict COVID-19 is computer-based, and it uses Python to test and train on COVID datasets. The object-oriented programming language "PYTHON" is used for a variety of activities it also includes automated machine learning methods for data mining.

Procedure

We are building a ML-based methodology that includes the four stages listed below.

Step 1: Considering the training-testing principle, construct a multi-class classification model.

The COVID-19 dataset's parameters for date, time, and state was obtained from Kaggle, and training and testing were undertaken at 80% and 20%, correspondingly.

Step2: Feature extraction.

To keep the model's complexity modest, we solely selected the much more critical features before initiating the prototype process. "Feature extraction" refers to processes that limit the quantity of data that must be processed while still precisely and thoroughly characterising the initial data set by picking certain variables and/or combining them to form features. The goal of extracting features is to reduce the amount of characteristics in a dataset by producing new ones from the existing ones.

Step 3: Training and testing by multi-classification

The data set is then modelled using NB, DT, RF, SVM, GB, and XGBoost ML techniques with 80% training and 20% testing.

ANALYSIS AND DISCUSSION

Once results are obtained and recorded, analysis of them becomes a crucial and attention-grabbing task. As a result, each outcome is thoroughly examined one by one. Results with a higher number of inaccurate data are ignored, and classification is again applied another time to those datasets after modification. In the results, it is also examined for true positives and false negatives. After a comprehensive examination of the findings, the results of the work are evaluated.

RESULTS AND DISCUSSION

Six approaches which have been demonstrated to be the most appropriate for the classification challenge to predict COVID-19 are:

- RF (Random-Forests)
- DT (Decision-Tree)
- SVM (Support Vector Machine)
- NB (NaïveBayes)
- GB (Gradient-Boosting)
- XGB (XGBoost)

After splitting dataset with training, 80%, and dataset 20% for test different ML models are applied to measure the accuracy for each three cases including Confirmed cases, Cured cases and for death cases.

Confirmed cases

For confirmed, cured and death cases almost all attributes of the dataset such as Date, Time, State or Union-Territory, Confirmed-Indian-National, Confirmed-Foreign-National, Cured Cases, Deaths Cases and Confirmed Cases are used. Table 2 shows Performance of Algorithms for Confirmed cases.The Performance of each algorithm is shown in table 2, which indicates the total count of correctly classified instances of class in percentage for the confirmed cases.

Table 2.
Performance for Confirmed cases

Classifier	Accuracy	Precision	Recall	F-Measure
GB	0.90	0.64	0.66	0.64
XGBoost	0.89	0.58	0.61	0.59

DT	0.89	0.58	0.59	0.58
RF	0.68	0.33	0.36	0.34
SVM	0.76	0.45	0.47	0.44
NB	0.89	0.58	0.60	0.59

Cured cases:

The Performance of each algorithm for cured cases is shown in table 3, which indicates the total count of correctly classified instances of class in percentage for the confirmed cases.

Table 3.
Performance for Cured cases

Classifier	Accuracy	Precision	Recall	F-Measure
GB	0.90	0.60	0.59	0.59
XGBoost	0.89	0.69	0.71	0.69
DT	0.83	0.78	0.73	0.73
RF	0.86	0.78	0.66	0.70
SVM	0.69	0.24	0.15	0.17
NB	0.24	0.40	0.30	0.33

Death cases

The performance of each algorithm for death is shown in table 4, which indicates the total count of correctly classified instances of class in percentage for the confirmed cases.

Table 4.
Performance for Death cases

Classifier	Accuracy	Precision	Recall	F-Measure
GB	0.92	0.64	0.50	0.55
XGBoost	0.88	0.69	0.66	0.69
DT	0.83	0.41	0.40	0.40
RF	0.86	0.71	0.61	0.63
SVM	0.74	0.42	0.34	0.35
NB	0.49	0.27	0.27	0.23

The overall result for confirmed, cured and death is shown in table 5.

Table 5.
Accuracy for all three cases

Model	Confirmed Accuracy	Cured accuracy	Death accuracy
GB	90%	90%	92%
XGBoost	89%	89%	88%
DT	89%	83%	83%
RF	68%	86%	86%
SVM	76%	69%	74%
NB	89%	49%	24%

The overall results of different models for different cases are mention in above table 5 and result shows that the Gradient Boosting Classifier gives more accurate result in each case and these results are graphically represented in figure 2.

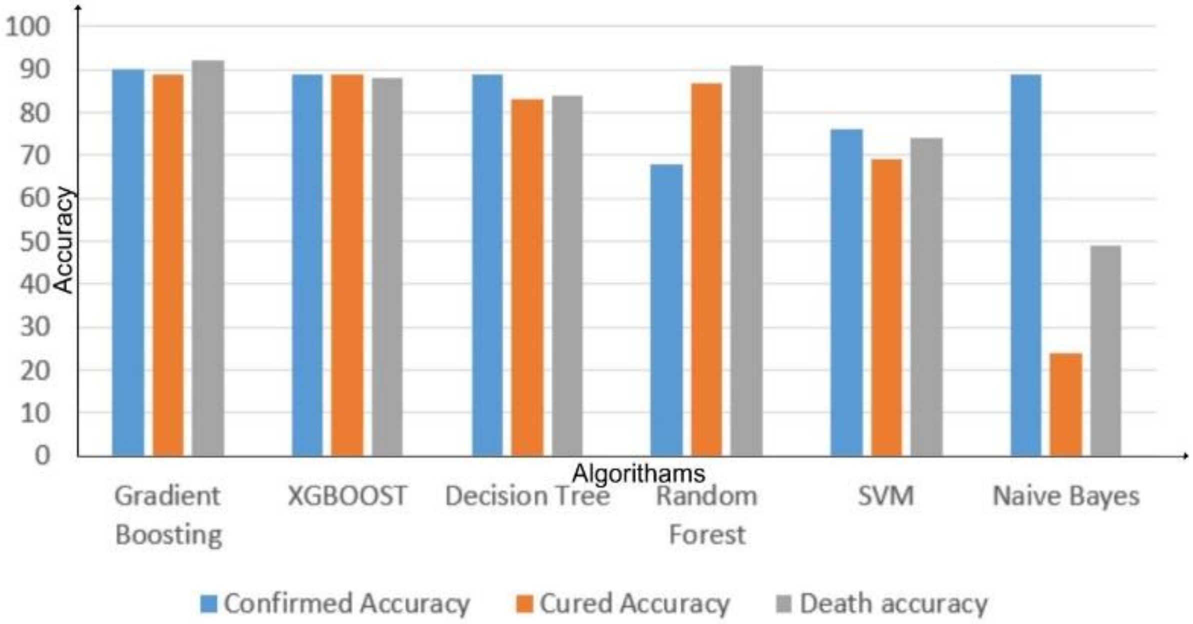

Figure 2.
Comparison for all three cases

CONCLUSION

There are numerous techniques and models for COVID-19 prediction, such as graph plotting and R-programming models, but these are time consuming and extensive. In this study, a comprehensive literature evaluation was done to determine the optimal technique to utilize for COVID-19 patient prediction. The primary objective of this research is to develop an efficient Machine learning model that will generate more accurate prediction results in a minimum amount of time while using the least amount of resources. For improvement of the accuracy, researcher employed several machine learning models such as RF, DT, SVM, NB, GB, XGBoost on numerical data sets. In order to test the accuracy of machine learning models, each algorithm is trained with sample sets including varying quantities of patient records. The performance of the trained algorithms was assessed using an accuracy performance indicator. After data analyses we come to know GB is out performed than XGBoost, RF, DT, SVM, NB. This research work determines Covd-19 cases for future. The proposed system will be implemented in the following two steps including preprocessing and training testing. After training the models with 80percent data from all above-mentioned models the Gradient Boosting Classifier give more accurate result with accuracy 90%, 90% and 92% correspondingly for confirmed cases, Cured, and Death cases. The next algorithms is XGBoost which give second more accurate result with accuracies of 89%, 89% and 88%. The third one is Decision Tree whose accuracies are 89%, 83% and 83%. The Random Forest is on fourth number with 68%, 86% and 86% and the fifth one is Support Vector Machine with 76%, 69% and 73% accuracies. The sixth and final is Naïve Bayes with accuracy 89%, 49% and 24% respectively.

FUTURE WORK

Healthcare offers a lot of promise for ML. Future research should concentrate on tailored and collective techniques that can address unusual problems more rapidly and efficiently than existing techniques. It is also conceivable to create an AI-based application that recognizes and classifies ailments using a broad range of sensors and parameters. It is possible to enlarge the purview of the suggested study to include the diagnosis of a variety of illnesses, including diabetes, heart disease, lung infections, biliary disorders, and more. A system that can anticipate recognize the danger of the breakout of novel ailments that might harm civilization by accounting for both socioeconomic and cultural aspects can be constructed, since healthcare prediction is a crucial subject for the future. For greater performance, one can also employ different classifier or a hybrid approach.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the supervisors and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., & Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimedia Tools and Applications*, *80*(8), 11943–11957. https://doi.org/10.1007/s11042-020-10340-7

Arslan, H. (2021). *Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data*. 20. https://doi.org/10.3390/proceedings2021074020

Azarafza, M., Azarafza, M., & Tanha, J. (2020). COVID-19 Infection forecasting based on deep learning in Iran. *MedRxiv*, 1–7.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, 105340. https://doi.org/10.1016/j.dib.2020.105340

Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Machine learning for coronavirus covid-19 detection from chest x-rays. *Procedia Computer Science*, *176*, 2212–2221. https://doi.org/10.1016/J.PROCS.2020.09.258

Daniyal, M., Ogundokun, R. O., Abid, K., Khan, M. D., & Ogundokun, O. E. (2020). Predictive modeling of COVID-19 death cases in Pakistan. *Infectious Disease Modelling*, *5*, 897–904. https://doi.org/10.1016/j.idm.2020.10.011

Fayyoumi, E., Idwan, S., & Aboshindi, H. (2020). Machine learning and statistical modelling for prediction of Novel COVID-19 patients case study: Jordan. *International Journal of Advanced Computer Science and Applications*, *11*(5), 122–126. https://doi.org/10.14569/IJACSA.2020.0110518

Gothai, E., Thamilselvan, R., Rajalaxmi, R. R., Sadana, R. M., Ragavi, A., & Sakthivel, R. (2021). Prediction of COVID-19 growth and trend using machine learning approach. *Materials Today: Proceedings*, 1–11. https://doi.org/10.1016/j.matpr.2021.04.051

Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*, *4*(2), 116–123. https://doi.org/10.26599/BDMA.2020.9020016

Holmes, K. V. (2003). SARS-Associated Coronavirus. *New England Journal of Medicine*, *348*(20), 1948–1951. https://doi.org/10.1056/nejmp030078

Huang, C.-J., Chen, Y.-H., Ma, Y., & Kuo, P.-H. (2020). Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China. *MedRxiv*, 2020.03.23.20041608.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology (Singapore)*, *12*(3), 731–739. https://doi.org/10.1007/s41870-020-00495-9

Kumari, R., Kumar, S., Poonia, R. C., Singh, V., Raja, L., Bhatnagar, V., & Agarwal, P. (2021). Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. *Big Data Mining and Analytics*, *4*(2), 65–75. https://doi.org/10.26599/BDMA.2020.9020013

Lasya, K. L., Lahari, D., Akarsha, R., Lavanya, A., Prakash, K. B., & Tran, D. T. (2022). Analysis and Prediction of COVID-19 datasets using Machine Learning Algorithms. *2022 1st International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2022*, *8*(5), 3–8. https://doi.org/10.1109/ICEEICT53079.2022.9768598

Mandayam, A. U., Rakshith, A. C., Siddesha, S., & Niranjan, S. K. (2020). Prediction of Covid-19 pandemic based on Regression. *Proceedings - 2020 5th International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2020*, 1–5. https://doi.org/10.1109/ICRCICN50933.2020.9296175

Mary, L. W., & Albert Antony Raj, S. (2021). Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) - A Comparative Analysis. *Proceedings - 2nd International Conference on Smart Electronics and Communication, ICOSEC 2021*, *2*, 1607–1611. https://doi.org/10.1109/ICOSEC51865.2021.9591801

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, *2*(1), 1–13. https://doi.org/10.1007/s42979-020-00394-7

Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19 Volume 1: Computational Perspectives*, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health*, *20*(November 2020), 100178. https://doi.org/10.1016/j.smhl.2020.100178

Rochmawati, N., Hidayati, H. B., Yamasari, Y., Yustanti, W., Rakhmawati, L., Tjahyaningtijas, H. P. A., & Anistyasari, Y. (2020). Covid Symptom Severity Using Decision Tree. *Proceeding - 2020 3rd International Conference on Vocational Education and Electrical Engineering: Strengthening the Framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering, ICVEE 2020*. https://doi.org/10.1109/ICVEE50212.2020.9243246

Samuel, A. L. (1959). Some Studies in Machine Learning. *IBM Journal of Research and Development*, *3*(3), 210–229. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560

She, J., Jiang, J., Ye, L., Hu, L., Bai, C., & Song, Y. (2020). 2019 novel coronavirus of pneumonia in Wuhan, China: emerging attack and management strategies. *Clinical and Translational Medicine*, *9*(1). https://doi.org/10.1186/s40169-020-00271-z

Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, *34*(7), 959–972. https://doi.org/10.1007/s00477-020-01827-8

Sun, N. N., Yang, Y., Tang, L. L., Dai, Y. N., Gao, H. N., Pan, H. Y., & Ju, B. (2020). A prediction model based on machine learning for diagnosing the early COVID-19 patients. *MedRxiv*, 1–12.

Van Der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J. M., Wolthers, K. C., Wertheim-Van Dillen, P. M. E., Kaandorp, J., Spaargaren, J., & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature Medicine 2004 10:4*, *10*(4), 368–373. https://doi.org/10.1038/nm1024

Venkata Ramana, B., Babu, M. S. P., & Venkateswarlu, N. . (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems*, *3*(2), 101–114. https://doi.org/10.5121/ijdms.2011.3207

Wernick, M., Yang, Y., Brankov, J., Yourganov, G., & Strother, S. (2010). Machine learning in medical imaging. *IEEE Signal Processing Magazine*, *27*(4), 25–38. https://doi.org/10.1109/MSP.2010.936730

Wiens, J., & Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153. https://doi.org/10.1093/cid/cix731

Zhao, H., Li, Y., Chu, S., Zhao, S., & Liu, C. (2021). A COVID-19 prediction optimization algorithm based on real-time neural network training - Taking italy as an example. *Proceedings of IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers, IPEC 2021*, 345–348. https://doi.org/10.1109/IPEC51340.2021.9421142