A sustainable advanced artificial intelligence-based framework for analysis of COVID-19 spread

Misbah Ahmad\textsuperscript{1} · Imran Ahmed\textsuperscript{2} · Gwanggil Jeon\textsuperscript{3}

Received: 3 October 2021 / Accepted: 21 April 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

The idea of sustainability aims to provide a protected operating environment that supports without risking the capacity of coming generations and to satisfy their demands in the future. With the advent of artificial intelligence, big data, and the Internet of Things, there is a tremendous paradigm transformation in how environmental data are managed and handled for sustainable applications in smart cities and societies. The ongoing COVID-19 (Coronavirus Disease) pandemic maintains a mortifying impact on the world population’s health. A continuous rise in the number of positive cases produced much stress on governing organizations worldwide, and they are finding it challenging to handle the situation. Artificial Intelligence methods can be extended quite efficiently to monitor the disease, predict the pandemic’s growth, and outline policies and strategies to control its transmission or spread. The combination of healthcare, along with big data, and machine learning methods, can improve the quality of life by providing better care services and creating cost-effective systems. Researchers have been using these techniques to fight against the COVID-19 pandemic. This paper emphasizes on the analysis of different factors and symptoms and presents a sustainable framework to predict and detect COVID-19. Firstly, we have collected a data set having different symptoms information of COVID-19. Then, we have explored various machine learning algorithms or methods: including Logistic Regression, Naive Bayes, Decision Tree, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, and Support Vector Machine to predict and detect COVID-19 lab results, using different symptoms information. The model might help to predict and detect the long-term spread of a pandemic and implement advanced proactive measures. The findings show that the Logistic Regression and Support Vector Machine outperformed from other machine learning algorithms in terms of accuracy; algorithms exhibit 97.66\% and 98\% results, respectively.

Keywords Artificial intelligence · Big data · Machine learning · Internet of Things · Sustainable healthcare · COVID-19
Advancement in artificial intelligence, the Internet of Things (IoT), and big data transformed the management and handling of data in sustainable applications. However, a crucial difficulty in artificial intelligence, and big data analytics, is the collection of data and the proper utilization of that data by applying advanced intelligent methods. Currently, COVID-19, a deadly disease or virus that originated in Wuhan, China, in late December 2019, has infected millions of people in the world and become life-threatening for thousands of people.¹ The outbreak’s initial spread made thousands of deaths, as the pharmaceutical organizations could not handle many severely sick or infected patients. In March 2020, the WHO (World Health Organization) declared it a pandemic, as the number of cases increased significantly across the globe. This declaration also highlighted the increasing alarm of the dangerous transmission or spread and severity of the virus. The administrative organizations in various counties are executing prohibitions, transport limitations, social distancing, and improving health awareness (Ahmed et al., 2021) and (Ahmed et al., 2021). However, still, the virus persists in spreading very quickly; people diagnosed with the COVID-19 virus have a mild pulmonary failure, whereas some have severe pneumonia. Older individuals having critical health issues, like asthma, coronary artery disorder, chronic lung cancer, and liver or kidney diseases, are expected to suffer severe infections. In February 2022, 386,548,962 confirmed cases of COVID-19, including 5,705,754 deaths, were reported across the globe because of this pandemic.² The latest figure of confirmed COVID-19 cases and deaths has been depicted in Fig. 1 (adopted from 2).

In a rapidly growing pandemic situation, inappropriate predictions and analysis of the disease/ infection/ virus severity and the number of patients may lead to an ineffective distribution of medical resources. In addition, lack of pharmaceutical resources and mismanagement of resource allocation, particularly in developing countries, can also cause extra intense cases, leading to lower recovery rates. Therefore, researchers have been developing several outbreak prediction systems (Ahmed et al., 2022), for COVID-19 to obtain notified decisions and enforce relevant management efforts to cope with this situation. Among these methods, worldwide pandemic prediction simple epidemiological and statistical designs have been gained more recognition from researchers (Ardabili et al., 2020). However,

---
¹ https://covid19.who.int/.
² https://ourworldindata.org/.
these methods have revealed low accuracy for long-term prediction because of the high uncertainty, lack of primary data, and due to emergence of different variants of COVID-19. Although numerous efforts have been made in the literature to cope with this issue, current designs’ central robustness and generalization capabilities require improvement.

Advance machine learning has shown encouraging results in healthcare (Ahmed et al., 2021), and (Ahmed et al., 2022) through its decision making and data analyzing capability. It also provides efficient monitoring, analysis, and prediction systems that allow fast and effective analysis of the COVID-19 pandemic to decrease healthcare organizations’ burden. Prediction systems that integrate different features to determine the uncertainty of infection have been introduced, which might assist pharmaceutical staff globally in tracking subjects, particularly with inadequate healthcare resources. These systems utilize different features such as X-Ray images (Ahmed et al., 2020), Computer Tomography (CT) scans (Gozes et al., 2020), clinical symptoms (Tostmann et al., 2020), laboratory tests (Punn et al., 2020), and combination of these features (Mei et al., 2020).

As discussed that COVID-19 has affected millions of people, and people in developing countries are susceptible to its future outcomes. Therefore, it is necessary to develop a system that will help to analyze, detect and predict the pandemic. One solution to manage the current devastation is the diagnosis of disease using various advanced machine learning methods. This paper analyzed textual clinical data set using machine learning algorithms such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, and Support Vector Machine. Different algorithms are trained using different clinical attributes information. The machine learning based system predicts COVID-19 test results (positive or negative) and patients final status (active, recovered, expired) with high accuracy using different features: patient id, age in years, gender, flu, fever, sore throat, cough, breathing issue, headache, cardiovascular & hypertension, chronic lung disease, foreign travel history, and test specimen information. The central contribution of the work is provided as follows:

- An intelligent and sustainable artificial intelligence-based system is developed for the detection of COVID-19 lab results and patient’s final status.
- Various machine learning algorithms are explored and implemented using textual data set to predict and classify the target attributes of COVID-19.
- Finally, the results of different algorithms are compared in terms of accuracy.

The paper is organized as follows: Sect. 2 provides a summary of related work. In Sect. 3, the system developed for the prediction of COVID-19 is presented. Different machine learning methods are also elaborated in this Section. The details of the data set and experimental outcomes are explained in Sect. 4. The conclusion and future guidelines of the work are finally given in Sect. 5.

2 Related work

Advanced artificial intelligence (Ahmed and Jeon, 2021; Ahmed et al., 2021, 2020), and (Ahmed et al., 2021a) along with big data analytics (Ahmed et al., 2021b), and IoT (Ahmad et al., 2021, 2020; Ahmed et al., 2021), and (Ahmed et al., 2021) helps to understand variables, behaviors, and data trends utilizing a broad range of algorithms and techniques that are substantially better than human performance.
Initial investigations studied different techniques like statistical methods, data mining, and artificial intelligence-based approaches (machine learning), supported segmentation methods (parallel, classified processing, statistical rationalizing), and distribution of big data applications for various types of control studies like prediction and risk (Khanday et al., 2020). This section presented a review of various artificial intelligence (machine learning-based algorithms) studied by researchers for analysis and detection of disease. In (Ardabili et al., 2020), authors provided a comparative analysis of soft computing and machine learning paradigms to estimate the outbreak.

Punn et al., (2020) employed machine and deep learning to identify growth and divine the future measure of the pandemic. Ahmed et al., (2021b), designed a framework applying big data analytics and IoT for the investigation and prediction of COVID-19. Authors performed various types of analysis, using different pandemic symptoms. In Khanday et al., (2020), researchers analyzed clinical data (textual) into four separate categories by using conventional and ensemble machine learning methods. Authors in Pashazadeh and Navimipour (2018) presented a broad survey on big data tools utilized in various healthcare practices using machine learning algorithms. Authors in Prakash et al., (2020) applied machine learning for the forecast and evaluation of COVID-19. Behnam and Jahanmahnin (2021), presented a data analytics system for predicting the COVID-19 virus. Authors compiled a data set that mainly contains the number of recovered, confirmed, and death cases reported every day using machine learning algorithms. Khakharia et al., (2021), produced a prediction method for the COVID-19 outbreak; the system is developed for the top ten extremely dense populated regions. Applying different machine learning algorithms, the models predict the number of newly reported cases expected to rise for five consecutive days.

Yan et al., (2020) applied machine learning to present a prognostic prediction method for the death risk of people affected, utilizing data of 29 subjects collected from Tongji Hospital in Wuhan, China. Jiang et al., (2020) added a machine learning method that predicts COVID-19 infected patients. The system produced 80% accuracy on sample data of 53 patients used for training. Rao and Vazquez (2020) produced a method based on artificial intelligence for the identification of subjects with COVID-19 using a mobile phone. Finally, authors in Yan et al., (2020) presented a modeling method to assess high-risk patients in the early stage using machine learning with three clinical features. Different researchers have made various research work on the prediction of COVID-19 (Peng et al., 2020). The study mainly focused on a conceptual method for different applications, using data mining methods (Peng et al., 2020).

Li et al., (2021) designed a regression model for the prediction of pandemic’s spread. Pinter et al., (2020) added a hybrid model for the outbreak prediction in Hungary. The model made no supposition on the transmission of the infection and pandemic; rather, it divines the time series analysis of the infected and fatality cases. Ahmad et al., (2020), offered a comprehensive study based on machine learning techniques used for investigation of the pandemic. Authors analyzed the difficulties and presented suggestions to the experts that might improve methods used to predict positive cases.

Rustam et al., (2020), described the potential of machine learning to find the number of predicted COVID-19 patients. In particular, four traditional prediction models have been practiced. Roy et al., (2020), performed research identifying the pandemic impacts worldwide utilizing the machine learning algorithms. The authors suggested an additive regression design that specialists normally evaluate with field knowledge about the time series. Ahmed and Jeon (2021) studies machine learning method for genome sequence analysis of the COVID-19 infection.
It is concluded, from the above discussion, that various machine learning methods (Ahmed et al., 2017, 2019; Ahmed and Carter, 2012; Ullah et al., 2019; Ahmed et al., 2019), and (Ahmad et al., 2019) have been used by researchers that helped in analyzing, detecting, and predicting the COVID-19 pandemic. However, it still needs many research efforts in order to enhance the performance. Therefore, we practiced machine learning algorithms for prediction of COVID-19 virus using a textual data set inspired by previous work.
3 Methodology

The presented machine learning-based system for predicting the COVID-19 textual data set is discussed in this section. Figure 2 explains the technical features and specifications of the system applied to predict the lab results of the COVID-19 virus. The data set is collected from different medical organizations, which are labeled with the help of medical experts. The raw collected clinical data are processed through the data processing layer.

After pre-processing, the data are split into training and testing and forwarded to the prediction stage, where different machine learning algorithms/techniques/methods are applied for training and testing purposes—the textual data set containing different attributes for the prediction of the lab results of COVID-19 virus. We used two target attributes "Lab results" and "Patient final status" as shown in Table 1 to develop a machine learning-based system to predict the lab results of virus, diagnosed as either positive or negative. Furthermore, another model is also trained for the prediction of final patient status as expired, alive, or recovered after diagnosing the virus, as shown in Fig. 2.

3.1 Data collection

The data set used in this paper is collected from different medical organizations, which contains data of 25,000 patients having symptoms of coronavirus (Ahmed et al., 2021b). Data consists of different attributes, namely patient id, age in years, gender, flu, fever, sore throat, cough, breathing issue, headache, cardiovascular, hypertension, chronic lung disease, foreign travel history, and specimen information. In addition, two attributes are defined as target attributes, namely Lab results and final patient status. The details of the collected data set have been shown in Table 1.

3.2 Data pre-processing

The data obtained from the data acquisition or previous step is in raw form, as it contains some inconsistent data and incomplete parameters. Thus, the data are processed in the pre-processing phase. The data cleansing has been performed where missing values are replaced using mean or mode, and outliers are removed. Further, the useful attributes are filtered and converted into binary form, e.g., flu symptom yes, and no is converted into 0, and 1 form, also depicted in Table 1.

3.3 Machine learning-based classification

Different machine learning algorithms have been practiced for prediction and classification purposes. The machine learning-based system first classifies the patient’s lab results as positive or negative on the basis of symptoms and then detects the patient’s final status, as shown in Fig. 2. Various machine learning algorithms are being utilized by researchers for classification purposes. Machine learning algorithms like Logistic Regression, Naive Bayes, Decision Tree, Random Forest Classifier, Extreme Gradient Boost, K-Nearest
Table 1  Detailed description of different attributes of the clinical data set used for experimentation

| S. no. | Attribute                        | Attribute code    | Attribute description                             |
|--------|----------------------------------|-------------------|--------------------------------------------------|
| 1      | Patient id                       | Patient number    | In numbers                                       |
| 2      | Age                              | Age               | Age in years                                     |
| 3      | Gender                           | Gender            | Male = 0, Female = 1                             |
| 4      | Is patient symptomatic?          | IspSymptamatic    | Yes = 1, No = 0                                  |
| 5      | Flu                              | Flu               | Yes = 1, No = 0                                  |
| 6      | Fever                            | Fever             | Yes = 1, No = 0                                  |
| 7      | Sore throat                      | SoreThroat        | Yes = 1, No = 0                                  |
| 8      | Cough                            | Cough             | Yes = 1, No = 0                                  |
| 9      | Breathing issue                  | Breathingissue    | Yes = 1, No = 0                                  |
| 10     | Headache                         | Headache          | Yes = 1, No = 0                                  |
| 11     | Cardiovascular & hypertension    | Cardiovascular and hypertension | Yes = 1, No = 0 |
| 12     | Chronic lung disease             | Chroniclung       | Yes = 1, No = 0                                  |
| 13     | Foreign travel history           | ForeignTravel History | Yes = 1, No = 0 |
| 14     | Test specimen information        | Specimen information | Nasopharyngeal swab = 0, Oropharyngeal swab=1 |
| 15     | Lab results                      | LabResults        | Positive = 1, Negative = 0                       |
| 16     | Patient final status             | PatientFinal Status | Expired = 0, Active = 1, Recovered = 2          |
Neighbour, Support Vector Machine are practiced in this work. The details of each algorithm are provided as follows.

### 3.3.1 Logistic Regression

It is a supervised algorithm utilized for the prediction of target variable possibility. It has two variables, the dependent and target, which means there would be two viable classes. The binary coded dependent variable is set as either 1 (yes/success) or 0 (no/failure). In our work, features shown in Table 1 are supplied as an input. The algorithm usually determines the class association probability. It defines a linear relationship between the dependent and independent variables. The relationship between the two factors is provided as;

\[
y = \frac{e(\beta_0 + \beta_1x)}{1 + e^{(\beta_0 + \beta_1x)}}
\]

In the above equation, \( y \) is the predicted output, and dependent variable, \( \beta \), are the weights or coefficient values, also known as the model parameters, \( x \) is the input values, and \( e \) represents an error term. The input values \( x \) are combined linearly using \( \beta \), here \( \beta_0 \) is the intercept term and \( \beta_1 \) is the coefficient for the single input value (\( x \)). In this work, we have two classes or predicted output results, including lab results as positive and negative, and three classes for final patient status (active, expired, recovered), represented as \( y = 0, 1 \), and \( y = 0, 1, 2 \), respectively.

### 3.3.2 Naive Bayes

It is an efficient and simple algorithm based on the Bayes’ theorem application, which strongly implies that a certain attribute of the sample is independent of other attributes. In other words, each feature of a sample contributes independently to determine the probability of the classification of that sample by providing the highest probability category of the sample. This algorithm makes the computation process easy and fast while producing accurate results for huge amounts of data. Let \( c \) indicate the set of classes; in our case, we have two classes for lab results as positive and negative, and three classes for patient final status (active, expired, recovered), represented as \( c = 0, 1 \), and \( c = 0, 1, 2 \), respectively. Furthermore, \( N \) is the number of attributes. The posterior probability is given as;

\[
P(c|x) = \frac{P(x|c)P(c)}{P(x)}
\]

where \( P(x|c) \) defines the posterior probability, used for target class \( c \), given for \( x \) input values or attributes. \( P(c) \) is used for the prior class probability. \( P(x|c) \) represents the class likelihood. \( P(x) \) indicates the predictor prior probability.

### 3.3.3 Decision Tree

The decision tree is utilized for both classification and regression. Its popularity comes from its simplicity, efficiency, and the fact that it is easy to implement, interpret, and explain. It is similar to a flowchart that comprises the root node, branches, and leaf nodes. Every node represents the condition or test; the branches represent the outcomes, whereas the leaf node indicates the class label. The classification rule is then determined by the route from the root to the
leaf node. The decision tree is built through several steps. Splitting is the process of dividing the data set into subsets on a particular variable. Pruning refers to reducing the tree’s size, obtained by converting branch nodes into leaf nodes. Pruning helps in avoiding over-fitting. First, the best attribute is chosen through Attribute Selection Measures (entropy) for splitting the records. Then that attribute is made the decision node for breaking the data set further into smaller subsets. The entropy is computed when the data are split into feature \( X \); this process is repeated recursively for each child until one of the following conditions reaches: All tuples belong to the same class, or no attribute remains. The information gain is utilized by splitting the data and applying entropy. It is estimated as the decrease in entropy after the data set is divided into attributes:

\[
\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)
\] (3)

In Eq. 3, \( T \) represents the target variable, and \( X \) is the features to be split on entropy \((T, X)\).

### 3.3.4 Random Forest classifier

Random Forest is a popular machine learning algorithm that is supervised in nature and can also be applied for classification problems. The popularity of this technique comes from its computing efficiency, even with large data sets and high dimensionality. It is an ensemble of decision trees that are suggested by its name and works like a decision tree with a key difference; this method builds a forest of decision trees with attribute locations chosen randomly. At the time of training, multiple decision trees are constructed and considered before producing an output. This technique is based on the idea of more trees that will reach the right decision. It employs a voting system before deciding the class in case of classification, whereas the output’s mean of all decision trees is taken for regression problems. The feature importance values for all trees are calculated as follows;

\[
\text{RFf}_{i} = \frac{\sum_{j} \text{normf}_{i,j}}{\sum_{j \in \text{features}, k \in \text{all trees}} \text{normf}_{i,k}}
\] (4)

In the above equation, \( \text{normf} \) is the normalized importance, and \( f_{i} \) is the importance of feature \( f \) feature \( i \) in tree \( j \), \( \text{RFf}_{i} \) is the \( i \) feature importance that is estimated for all trees in the model \( \text{normf}_{i,j} \).

### 3.3.5 Extreme Gradient Boost

Extreme Gradient Boosting is the most widely used machine learning algorithm. It can be used for supervised learning applications. It is built on a gradient boosting framework and based on function approximation. It optimizes specific loss functions utilizing several regularization techniques. One of this algorithm’s main features is its scalability, which forces fast learning throughout distributed, parallel computing and gives effective memory usage. For an assigned data set with \( n \) samples and \( m \) number features \( D = \{(x_{i}, y_{i})\} \), where \(|D| = n\), and \( x_{i} \in \mathbb{R}^{m} \), \( y_{i} \in \mathbb{R} \), \( K \) number of trees, the output is predicted as;

\[
y_{i} = \sum_{k=1}^{K} f_{k}(x_{i}), f_{k} \in F
\] (5)

The objective function at iteration \( t \) is provided as;
In the above equation, \( y_i \) is the real value or label known from the training data set, \( y'_{i(t)} \) is utilized for the prediction of the \( i \)th sample at the \( t \)th iteration, and \( f_t \) is employed to minimize the objective function.

### 3.3.6 K-Nearest Neighbour

It is a simple algorithm that assembles all cases and classifies them using similarity measures (for example, distance functions). It has been utilized in statistical evaluation and pattern recognition. It is a nonparametric method and can be applied for both regression and classification. The algorithm is based on feature similarity. This means that the classification of new data points will be based on their similarity to the training set’s data points. An object is allocated to the class most prevalent among its \( k \) nearest neighbors. The value of \( k \) determines the number of neighbors chosen initially. Different distance functions have been used, such as;

\[
D_{\text{Euclidean}} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}
\]

\[
D_{\text{Manhattan}} = \sum_{i=1}^{k} |x_i - y_i|
\]

\[
D_{\text{Minkowski}} = \left( \sum_{i=1}^{k} |x_i - y_i|^q \right)^{\frac{1}{q}}
\]

### 3.3.7 Support Vector Machine

It is a traditional and effective supervised machine learning technique employed for regression and classification problems. It defines a hyperplane, which separates the data into two categories in feature space. The greater the possibilities are of obtaining classification accuracy as more data points drop from the hyperplane accurately. The nearest data points are related to support vectors. If these support vectors are dropped, the state of the hyperplane changes; therefore, they are supposed to be the essential components of the data set. The gap between both sides is correlated with the margin. The aim is to choose a hyperplane with the biggest edge that lies between every training set point to accurately incorporate the new data points. The objective function is given as follows;

\[
L(w) = \sum_{i=1}^{n} \left[ \max(0, 1 - y_i[w^Tx_i + b]) + \lambda ||w||^2 \right]
\]

In the above equation, there are two terms; one is for regularization, and the other is for loss. The loss term is applied to penalize missclassifications, which calculates the error,
while the other one is utilized to evade over-fitting. The regularization coefficient is defined with $\lambda$.

4 Experimental results

To evaluate the performance of above discussed algorithms, we used different evaluation metrics as follows:

- True Positive (TP) defines correctly positive predicted classes.
- False Positive (FP) used for incorrectly positive predicted classes.
- True Negative (TN) defines correctly negative predicted classes.
- False Negative (FN) used incorrectly negative predicted classes.

Using the above matrices, we estimated the following performance measuring parameters; Accuracy is the ratio of the correct predictions and total input samples, evaluated as follows;

$$\text{Accuracy} = \frac{TP}{TP+TN+FP+FN}$$  \hspace{1cm} (11)

Precision is the estimated as follows;

$$\text{Precision} = \frac{TP}{TP+FP}$$  \hspace{1cm} (12)

Recall is provided as;

$$\text{Recall} = \frac{TP}{TP+FN}$$  \hspace{1cm} (13)

F1-Score ranges between [0, 1], and mean between recall and precision. It essentially defines how close the classifier is to accuracy (how many samples are accurately labeled) and how strong it is. The higher the F1-Score, the greater is the performance. Mathematically, estimated as;

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$  \hspace{1cm} (14)

True Positive Rate (TPR) is defined as;

$$\text{TPR} = \frac{TP}{TP + FN}$$  \hspace{1cm} (15)

True Negative Rate TNR is mathematically, evaluated as;

$$\text{TNR} = \frac{TN}{TN+FP}$$  \hspace{1cm} (16)

False Positive Rate FPR indicates the amount of negative samples that are mistakenly classified as true class, defined as;
Fig. 3  ROC curve plotted using TPR versus FPR

Fig. 4  Precision, Recall and F1-Score, of different machine learning algorithms
FPR and TPR have values in the range [0, 1]. The Receiver Operating Characteristic (ROC) Curve is plotted, as shown in Fig. 3 using the FPR values versus the TPR. We can observe from the Figure that all algorithms performed well and were nearly equal to 90%, but the performance of SVM is outperformed among all.

As we discussed, Precision specifies the number of true class predictions that actually exist in the true class, while Recall computes the number of true class predictions obtained from all true samples in the data set. F1-Score provides a score that assesses both the interests of Precision and Recall in unit number. Thus, we have also shown the Precision, Recall, and F1-Score results in Fig. 4; it can be seen that Logistic Regression and SVM values are higher than compared to other algorithms, which shows the robustness of both algorithms. The Recall, Precision, and F1-Score of Logistic Regression and SVM 98, 97 and 98%. The value of Extreme Gradient Boost is small as compared to other algorithms.

The accuracy comparison of various machine learning algorithms used in this work has been provided in Fig. 5, it can be seen that the accuracy of Logistic Regression,
Naive Bayes, Decision Tree Random Forest, and K-Nearest Neighbour is more than 96%. The accuracy of Extreme Gradient Boost is nearly equalled to 95% as compared to other algorithms; the SVM outperforms among all with an accuracy of 99%.

The comparison results of algorithms are shown in Table 2; it can be seen that the performance of Logistic Regression and SVM is good as compared to other machine learning algorithms.

5 Conclusion and future work

This paper highlights the importance of machine learning algorithms to analyze different symptoms and factors of the COVID-19 in order to predict and detect infected patients’ lab results and their final status. A clinical data set has been used for the analysis of different symptoms of COVID-19 disease. We have explored different machine learning algorithms such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, and Support Vector Machine to predict and detect COVID-19 on the basis of different symptoms. The system mainly applied supervised machine learning algorithms to analyze and predict the disease, which might help to predict the long-term transmission of an outbreak in order to implement advanced proactive measures. The findings reveal that the Logistic Regression and Support Vector Machine outperformed and archived 97.66% and 98% results in terms of accuracy. We intend to extend this work to other machine learning algorithms in the future. Also, the effectiveness of algorithms can be enhanced by increasing the amount of data.

References

Ahmad, M., Ahmed, I., & Jeon, G. (2021). An IoT-enabled real-time overhead view person detection system based on Cascade-RCNN and transfer learning. Journal of Real-Time Image Processing, 18(4), 1129.
Ahmad, M., Ahmed, I., Khan, F. A., Qayum, F., & Aljuaid, H. (2020). Convolutional neural network-based person tracking using overhead views. International Journal of Distributed Sensor Networks, 16(6), 155014720934738.
Ahmad, M., Ahmad, I., Ullah, K., Khan, I., Khattak, A., & Adnan, A. (2019). Person detection from overhead view: A survey. International Journal of Advanced Computer Science and Applications, 10(4), 567.
Ahmad, A., Garhwal, S., Ray, S. K., Kumar, G., Malebary, S. J., & Barukab, O. M. (2020). The number of confirmed cases of COVID-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-020-09472-8.
Ahmed, I & Carter, J.N. (2012). A robust person detector for overhead views, in Proceedings of the 21st international conference on pattern recognition (ICPR2012) (IEEE), pp. 1483–1486.
Ahmed, I., Ahmad, M., Adnan, A., Ahmad, A., & Khan, M. (2019). Person detector for different overhead views using machine learning. International Journal of Machine Learning and Cybernetics, 10(10), 2657.
Ahmed, I., Ahmad, M., Ahmad, A., & Jeon, G. (2021). IoT-based crowd monitoring system: Using SSD with transfer learning. Computers & Electrical Engineering, 93, 107226.
Ahmed, I., Ahmad, A., & Jeon, G. (2020). An IoT based deep learning framework for early assessment of COVID-19. IEEE Internet of Things Journal, 8, 15855–15862.
Ahmed, I., Ahmad, M., & Jeon, G. (2021). Social distance monitoring framework using deep learning architecture to control infection transmission of COVID-19 pandemic. Sustainable Cities and Society, 69, 102777.
Ahmed, I., Ahmad, M., Jeon, G., & Piccialli, F. (2021). A framework for pandemic prediction using big data analytics. Big Data Research, 25, 100190.
Ahmed, I., Ahmad, M., Khan, F. A., & Asif, M. (2020). Comparison of deep-learning-based segmentation models: Using top view person images. *IEEE Access*, 8, 136361.

Ahmed, I., Ahmad, A., Piccialli, F., Sangaiah, A. K., & Jeon, G. (2017). A robust features-based person tracker for overhead views in industrial environment. *IEEE Internet of Things Journal*, 5(3), 1598.

Ahmed, I., Ahmad, M., Rodrigues, J. J., & Jeon, G. (2021). Edge computing-based person detection system for top view surveillance: Using CenterNet with transfer learning. *Applied Soft Computing*, 107, 107489.

Ahmed, I., Ahmad, M., Rodrigues, J. J., Jeon, G., & Din, S. (2021). A deep learning-based social distance monitoring framework for COVID-19. *Sustainable Cities and Society*, 65, 102571.

Ahmed, I., Anisetti, M., & Jeon, G. (2021). An IoT-based human detection system for complex industrial environment with deep learning architectures and transfer learning. *International Journal of Intelligent Systems*. https://doi.org/10.1002/int.22472.

Ahmed, I., Camacho, D., Jeon, G., & Piccialli, F. (2022). Internet of health things driven deep learning-based system for non-invasive patient discomfort detection using time frame rules and pairwise keypoints distance feature. *Sustainable Cities and Society*, 79, 103672.

Ahmed, I., Din, S., Jeon, G., & Piccialli, F. (2019). Exploring deep learning models for overhead view multiple object detection. *IEEE Internet of Things Journal*, 7(7), 5737.

Ahmed, I., & Jeon, G. (2021). A real-time person tracking system based on SiamMask network for intelligent video surveillance. *Journal of Real-Time Image Processing*, 18(5), 1803.

Ahmed, I., & Jeon, G. (2021). Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. *Interdisciplinary Sciences: Computational Life Sciences*. https://doi.org/10.1007/s12539-021-00465-0.

Ahmed, I., Jeon, G., & Chehri, A. (2022). An IoT-enabled smart health care system for screening of COVID-19 with multi layers features fusion and selection. *Computing*, 10, 1–18.

Ahmed, I., Jeon, G., Chehri, A., & Hassan, M. M. (2021). Adapting Gaussian YOLOV3 with transfer learning for overhead view human detection in smart cities and societies. *Sustainable Cities and Society*, 70, 102908.

Ahmed, I., Jeon, G., & Piccialli, F. (2021). A deep-learning-based smart healthcare system for patient’s discomfort detection at the edge of Internet of Things. *IEEE Internet of Things Journal*, 8(13), 10318.

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., et al. (2020). COVID-19 outbreak prediction with machine learning. *Algorithms*, 13(10), 249.

Behnam, A., & Jahanmahin, R. (2021). A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran). *Modeling Earth Systems and Environment*. https://doi.org/10.1007/s40808-021-01086-8.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W., Bernheim, A & Siegel, E. (2020). Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis, arXiv preprint arXiv:2003.05037

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., et al. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua*, 63(1), 537.

Khakharia, A., Shah, V., Jain, S., Shah, J., Tiwari, A., Daphal, P., et al. (2021). Outbreak prediction of COVID-19 for dense and populated countries using machine learning. *Annals of Data Science*, 8(1), 1.

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. U. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, 12(3), 731.

Li, Y., Liang, M., Yin, X., Liu, X., Hao, M., Hu, Z., et al. (2021). COVID-19 epidemic outside China: 34 founders and exponential growth. *Journal of Investigative Medicine*, 69(1), 52.

Mei, X., Lee, H. C., Dao, K.-Y., Huang, M., Lin, B., Liu, C., et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, 26(8), 1224.

Pashazadeh, A., & Navimipour, N. J. (2018). Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. *Journal of biomedical informatics*, 82, 47.

Peng, L., Yang, W., Zhang, D., Zhuge, C & Hong, L. (2020). Epidemic analysis of COVID-19 in China by dynamical modelling. arXiv preprint arXiv:2002.06563

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary: A hybrid machine learning approach. *Mathematics*, 8(6), 890.

Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. (2020). Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms. *International Journal*, 8(5), 2199–2204.
Punn, N.S., Sonbhadra, S.K & Agarwal, S. (2020). COVID-19 epidemic analysis using machine learning and deep learning algorithms, MedRxiv.

Rao, A. S. S., & Vazquez, J. A. (2020). Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. *Infection Control & Hospital Epidemiology, 41*(7), 826.

Roy, A.N., Jose, J., Sunil, A., Gautam, N., Nathalia, D & Suresh, A. (2020). Prediction and spread visualization of COVID-19 pandemic using machine learning. Preprints retrieved from 10.20944/preprints202005.0147.v1

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access, 8*, 101489.

Tostmann, A., Bradley, J., Bousema, T., Yiek, W. K., Holwerda, M., Bleeker-Rovers, C., et al. (2020). Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. *Eurosurveillance, 25*(16), 2000508.

Ullah, K., Ahmed, I., Ahmad, M., Rahman, A. U., Nawaz, M., & Adnan, A. (2019). Rotation invariant person tracker using top view. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-019-01526-5.

Yan, L., Zhang, H.T., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Li, S & Zhang, M. et al. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan, MedRxiv.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

**Authors and Affiliations**

**Misbah Ahmad**¹ · **Imran Ahmed**² · **Gwanggil Jeon**³

Misbah Ahmad  
misbahahmad4872@gmail.com

Imran Ahmed  
imran.ahmed@aru.ac.uk

¹ Center of Excellence in Information Technology, Institute of Management Sciences, 1-A, Sector E-5, Phase VII, Peshawar, Hayatabad, Pakistan

² School of Computing and Information Science, Anglia Ruskin University, Cambridge East Road, Cambridge CB1 1PT, UK

³ Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea