Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 31 (2022) 100990

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Prediction of COVID-19 using long short-term memory by integrating principal component analysis and clustering techniques

Saratu Yusuf Ilu [a,*], Prasad Rajesh [a], Hassan Mohammed [b]

[a] African University of Science and Technology, Abuja, Nigeria
[b] Bayero University, Kano, Nigeria

ARTICLE INFO

Keywords:
COVID-19
LSTM
Machine learning
Classification and clustering

ABSTRACT

Severe acute respiratory syndrome coronavirus (SARS-COV) is a major family of viruses that cause infections in both animals and humans, including common cold, coronavirus disease (COVID-19), severe acute respiratory syndrome (SARS), and Middle East respiratory syndrome. This study primarily aims to predict the number of COVID-19 positive cases in 36 states of Nigeria using a long short-term memory (LSTM) algorithm of deep learning. The proposed approach employs K-means clustering to detect outliers and principal component analysis (PCA) to select important features from the dataset. The LSTM was chosen because of its non-linear characteristics to handle the dataset. As COVID-19 cases follow non-linear characteristics, LSTM is the most suitable algorithm for predicting their numbers. For comparison, several types of machine learning algorithms, such as naive Bayes, XG-boost, and SVM, were employed. After the comparison, LSTM was observed to be superior among all algorithms.

1. Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV-2) causes the coronavirus disease (COVID-19), which first emerged in the Chinese city of Wuhan in late 2019 and caused significant public health and socioeconomic unrest [1]. Difficulty in breathing, fever, coughing, and sore throat are some of the basic symptoms of COVID-19. COVID-19 spreads among people in different ways: people in close contact with each other. For example, at a conversational distance, the virus can spread from an infected person's mouth or nose via droplets produced while coughing, sneezing, speaking, singing, or breathing. The virus can also spread in poorly ventilated and/or crowded indoor settings, where people tend to spend longer periods of time. People may also become infected by touching their eyes, nose, or mouth after touching surfaces that have been contaminated by the virus [2]. The disease emerged from a traditional seafood market in late December 2019 in Wuhan, China. Furthermore, it gradually spread to many other countries [3]. Initially, the infected patients were diagnosed with pneumonia. Consequently, in early January 2020, 41 patients were tested positive for COVID-19. Similarly, 25 jurisdictions in China reported 571 COVID-19 cases on January 22, 2020. Furthermore, China certified 7734 COVID-19 instances. Ninety more instances were verified in 13 countries, as of January 30, 2020, including Canada, the United Arab Emirates, Germany, and the United States.

On May 27, 2020, a total of 5,656,615 COVID-19 cases were reported worldwide. Several preventive measures were adopted by the Chinese government to curtail the outbreak due to the rising number of cases. In early 2020, the authorities of Wuhan, China, shut down all modes of transportation, including road, rail, and air and imposed quarantine on all other Chinese cities [4].

To prevent the global spread of COVID-19, government authorities employed a variety of measures, including suppression and a combination of mitigation strategies. Some of these strategies include social isolation, mandatory usage of face masks, frequent hand washing, ceasing any public or private activities except for vital services, and home isolation for positive and suspected COVID-19 cases. Additionally, to further reduce the spread of the virus, detected cases were instantly placed in isolation for immediate treatment, and many countries monitored individuals who have may have been in contact with infected individuals [3–5].

The first Nigerian COVID-19 case was declared on February 27, 2020. In addition, on May 27, 2020, a total of 8733 new cases of COVID-19 were recorded, with 5978 active cases, 254 fatalities, and 2501 discharged cases. The Lagos state of COVID-19 epicenter, reported 4012

* Corresponding author.
E-mail addresses: syilu@aust.edu.ng (S.Y. Ilu), rprasad@aust.edu.ng (P. Rajesh), mhassan.se@buk.edu.ng (H. Mohammed).

https://doi.org/10.1016/j.imu.2022.100990
Received 9 March 2022; Received in revised form 1 June 2022; Accepted 1 June 2022
Available online 3 June 2022

confirmed cases, 3220 active cases, 47 fatalities, and 745 discharged cases on the same day. It is noteworthy that, in a population of over 200 million people, only 48,544 samples were tested for COVID-19 on May 27, 2020. Consequently, many COVID-19 cases in the general population went unnoticed, causing an underreporting of the disease's actual prevalence in the country. Therefore, more cases of COVID-19 are expected to be detected using an enhanced testing rate, which will provide clearer insight into the gravity of the pandemic [5].

ML algorithms are used to improve patients' health by enhancing the prognosis, monitoring, diagnosis, and therapy management. Since the outbreak of the COVID-19 pandemic, researchers have demonstrated great interest in employing ML approaches to mitigate it. The authors of [6] built a model that estimated COVID-19 transmission in the 10 most afflicted states of India for 30 days, using a system modeling and identification method. Their results showed that their model can accurately capture variations in COVID-19 cases, which exhibit a rapid increase in cumulative cases and death rates. However, their prediction depends on certain system parameters and can differ based on factors, such as social distancing, vaccines, and lockdown.

In the Lagos state of Nigeria, a mathematical model was employed to develop a forecasting technique for aggregate cases and COVID-19 positive. This model depicts the consequences of preventative strategies, including mandatory face mask usage, social distancing, case discovery by contact-tracking, and follow-up testing. The authors of [7] studied 151 published studies and analyzed COVID-19 data using the XGBoost model by incorporating data from 413 patients. Based on clinical data variables, they divided the patients into subgroups and used the model to identify them as patients with COVID-19 or influenza. The model used patient symptoms and regular test results to predict the presence of the virus. Their findings suggest that computational tools trained on a large clinical dataset might lead to more precise COVID-19 diagnosis models, thereby reducing the impact of testing gaps. According to the findings, COVID-19 patients were identified from influenza patients with a sensitivity of 92.5% and a specificity of 97.9%.

The COVID-19 dataset exhibited non-linear characteristics. A dataset is non-linear if it is not linearly separable. As shown in Fig. 1, a scatter diagram was plotted, taking the number of weeks on the X-axis (for a period of 20 weeks) and the number of cases on the Y-axis. This shows that there is no clear-cut separation of cases when the number of weeks increases. However, the number of cases did not follow a definite trend. Owing to these characteristics, the aforementioned algorithms are not sufficiently accurate. This study provides a deep learning-based strategy for predicting and evaluating COVID-19 positive instances that incorporate principal component analysis (PCA), K-means clustering (KMC), and RNN-based long short-term memory (LSTM) deep learning algorithms. The LSTM algorithm is the most suitable when the dataset is non-linear. The dataset was studied using XG-Boost, SVM, and the naive Bayes model, among other ML approaches.

The remainder of this paper is organized as follows. Section 2 describes the related work; Section 3 describes the system architecture, methodologies, and techniques used in this research; and Section 4 describes the system architecture. Section 5 summarizes the findings and conclusions, and finally recommendations for future studies are discussed in Section 6.

2. Related work

Numerous studies have been published to predict COVID-19 instances, including the following. Researchers of [8] developed a machine learning (ML) technique for COVID-19 patients utilizing the XGBoost algorithm and deep neural network based on routine blood tests, age, and sex. They used data on 5333 patients suffering from various bacterial and viral infections. A total of 160 patients tested positive and were admitted to the University Medical Center, Ljubljana. The model produced a 97% area under the receiver operating characteristic curve, 81.9% sensitivity, and 97.9% specificity. The international normalized ratio (INR), albumin level, eosinophil count, and percentage of prothrombin activity are the most critical features of this model for forecasting the virus. The most important features that distinguish bacterial infections from COVID-19 are urea, erythrocyte count, hemoglobin, leukocyte count, and hematocrit.

A ML model was adopted utilizing the XGBoost algorithm to forecast positive and negative COVID-19 patients using 20 laboratory tests. Similarly, the US Department of Veterans Affairs examined data on 75991 cases, out of which 7335 cases were found to be positive [8]. The accuracy of the findings was 86.4%, with a specificity of 86.8% and a sensitivity of 82.4%. They discovered that their model, which is based on a comprehensive collection of independent indicators, provides a complementary technique for determining SARS-CoV-2 status, assisting in recognizing results from other tests and even detecting confirmed cases missed by molecular testing.

Researchers of [9] used the random forest algorithm to develop an assistant discrimination tool for identifying COVID-19 patients using 11 blood indices. A total of 253 samples of data were collected, of which

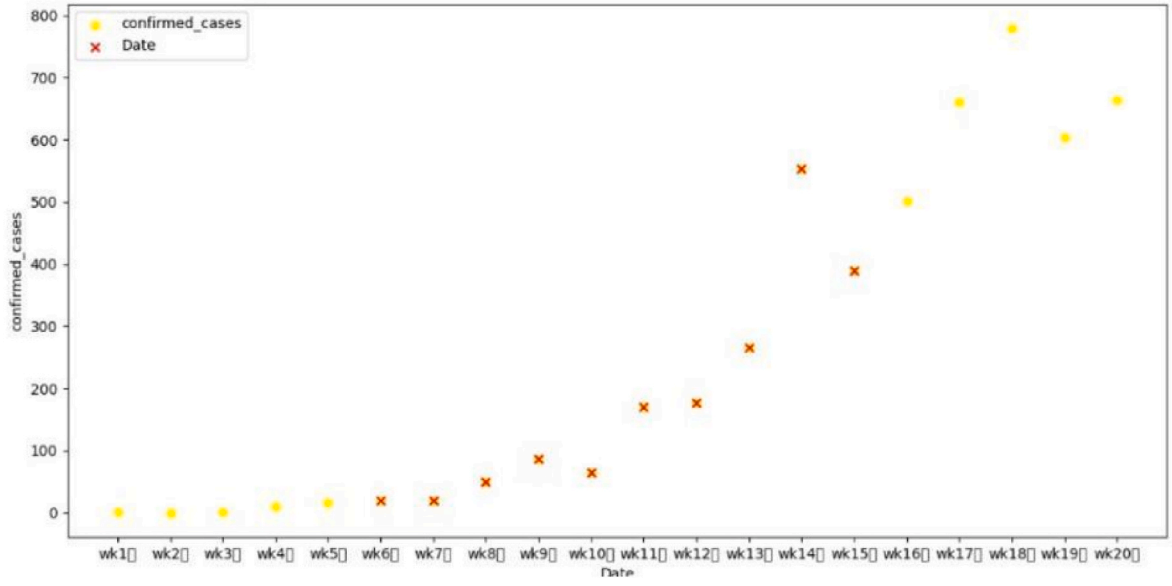

Fig. 1. Scatter diagram for COVID-19 cases and dates.

169 were suspected to be COVID-19 positive in China. Their model correctly identified COVID-19 patients with related symptoms, with a cross-validation set accuracy of 0.9795 and a test set accuracy of 0.9697. The model outperformed the baseline models, with a sensitivity of 0.9512, specificity of 0.9697, and accuracy of 0.9595. They realized that their model would be crucial if used for many early screenings on COVID-19 suspected patients. However, this approach should be tested using additional clinical practices.

Using petrochemical results from standard blood tests, two ML classification algorithms for COVID-19 prediction were developed [10]. The dataset was obtained from 279 COVID-19 suspected patients at San Raffaele Hospital, Italy. Among the patients, 177 tested positive, while the other 102 tested negative. Their model revealed the possibility and utility of using blood test analysis and ML, instead of RT-PCR, to discover COVID-19 positive patients. Several ML approaches were used by the model, such as ET, RF, KNN, and SVM among others, out of which the RF classifier achieved the best performance, with 84% area under the curve (AUC), 82% accuracy, 92% sensitivity, and 65% specificity, which was the fundamental contribution of the model.

Researchers of [11] developed a ML COVID-19 prediction model using a random forest algorithm based on routinely available laboratory values. They used information from 1528 patients where 65 were confirmed positive. The model had an accuracy of 81%, an AUC of 0.74, 60% sensitivity, and 82% specificity. They discovered that they could predict the findings of SARS-CoV-2 reverse transcription polymerase chain reaction (RT-PCR) with a reasonable degree of accuracy by using common blood parameters. They also found that, when predicting COVID-19, the most important features to consider were red blood cell count and hemoglobin level.

A logistic regression strategy was used to construct a ML model using two variables: the patient's blood count and sex [12]. Data were obtained from the blood count COVID-19 test results of emergency department patients who were above 18 years of age. At Stanford Health Care, the model was trained on 357 patients who tested negative and 33 patients who tested positive for COVID-19. Three blood count components were considered as features: absolute lymphocyte count, hematocrit, absolute neutrophil count, and male sex. Their findings revealed that a decision support tool has a high negative predictive value for COVID-19 RT-PCR results in various northwestern states of the US and South Korean populations. Their approach was able to achieve a sensitivity between 86% and 93%, specificity between 35% and 55%, and C-statistic of 78%.

The researchers used nine simple survey questions to develop an approach that estimates the likelihood of a person testing positive for the virus [13]. They used a dataset obtained from a survey of Israeli residents from different cities. Datasets from 43,752 adults were used, of which 498 were confirmed positive. The model was trained using a logistic regression technique, which yielded an AUROC of 0.737 and an AuPR of 0.144. Their model is good; however, it depends on self-reports from interested participants, which may lead to bias and non-objective assessment of symptoms.

Using a binary logistic regression technique, researchers of [14] developed a scoring model that predicts the chance of COVID-19 positivity. From three hospitals in France, they included 400 patients with clinical COVID-19 infection, of which 258 were positive. Their findings performed well in the validation cohort, with an AUC of 88.9%, a sensitivity of 80.3%, and a positive predictive value of 92.3%. Lymphocytes, basophils, neutrophils, and eosinophils were strongly associated with COVID-19 diagnosis.

The number of COVID-19 cases in India's 32 states has been forecasted using a variety of recurrent neural networks (RNNs) with long short-term memory (LSTM) [15]. To aid the identification of COVID-19 hotspots, states were grouped according to their case counts and daily growth rates. The researchers developed a website for scholars, authorities, and planners in which their model's state-by-state projections are updated. Their findings indicate that the model's accuracy is very high for short-term forecasting, with less than 3% error for daily and 8% for weekly forecasts.

Using an LSTM classification model, this paper proposes a ML technique that integrates PCA and KK-means clustering to forecast COVID-19 cases. To our knowledge, the approaches proposed in this study have not been adopted by researchers in Nigeria. This study makes significant contributions by predicting the number of positive cases, which enables governments to design preventive measures. The outcomes of this study will contribute to a better understanding of the rate of COVID-19 transmission in Nigeria, which will aid in the development of preventative measures.

3. System architecture

This section outlines various steps adopted for implementation of the COVID-19 prediction model.

3.1. Study site

All 36 states of Nigeria and the federal capital territory were selected for the study.

3.2. Dataset description

This study analyzed the Nigerian COVID-19 dataset retrieved from the Nigerian Center for Disease Control (NCDC). From February 29, 2020, when the first COVID-19 case was reported, to January 15, 2022, the dataset contains 254,124 confirmed weekly cases from 36 states and the federal capital area, all of which were tested for COVID-19. The dataset contains six attributes with a single target class (confirmed) and the following attributes: impacted states, confirmed (total/last week), recovery (total/last week), deaths (total/last week), and active testing (total/last week).

A statistical method for missing data, known as linear weighted moving average, was used to input missing values in each individual series to maintain the model's sequential learning capacity and prediction accuracy. The data were divided into training and testing groups using K-fold cross-validation. Table 1 represents a sample dataset used in this study.

3.3. Data preprocessing

Data preprocessing is a vital stage in the knowledge discovery process because good and accurate decisions are achieved from quality data. It can be described as the process of cleaning, selecting, transforming, extracting features, and normalizing data [16]. Real-world data are typically deficient, dirty, incomplete, and unreliable. Very good and reliable decision-making is achieved when data irregularities are detected and corrected at an early stage of data processing, which improves the efficiency and accuracy of the succeeding data mining process.

Data quality is crucial for disease prediction and diagnosis using ML. Inadequate data quality and efficiency may result in erroneous or inferior predictions r. To enhance the usefulness and applicability of our initial dataset for COVID-19 prediction, multiple preprocessing techniques were applied in conjunction with various tools available in the Python programming environment. Table 2 represents a sample of the dataset after preprocessing.

4. Proposed methodology

The proposed approach was designed and implemented using three ML algorithms: PCA, K-means, and LSTM models. PCA is used to filter out any unsuitable features that help improve the model's performance and reduce the training time and cost [17]. The K-means clustering algorithm then uses the PCA result, which helps remove outliers [18]. The

Table 1
Sample of the dataset before preprocessing from (https://covid19.ncdc.gov.ng/report/#!).

States	Confirmed		Recoveries		Deaths		Active cases	Testing	
	Total	Last week	Total	Last Week	Total	Last week		Total	Last week
Abia	2, 153	1	2, 118	6	34	0	1	51, 549	5, 782
Adamawa	1, 203	0	1, 103	0	32	0	68	30, 873	36
Akwa Ibom	4, 638	0	4, 562	0	44	0	32	57, 574	4, 397
Anambra	2, 825	0	2, 760	0	19	0	46	55, 303	216
Bauchi	1, 939	0	1, 882	0	24	0	33	37, 712	7
Bayelsa	1, 310	2	1, 277	0	28	0	5	37, 682	83
Benue	2, 129	0	1764	0	25	0	340	50, 361	210
Borno	1, 629	0	1, 580	0	44	0	5	28, 135	294
Cross River	805	20	760	7	25	0	20	18, 976	106

Table 2
Sample of dataset after preprocessing.

Date	Id	confirmed cases
January 02, 2021	46	9940
January 09, 2021	47	10300
January 16, 2021	48	11179
January 23, 2021	49	9676
January 30, 2021	50	8506
February 06, 2021	51	6606
February 13, 2021	52	5720
February 20, 2021	53	3583
February 27, 2021	54	2878
February 27, 2021	55	2122

proposed classification model for the COVID-19 dataset was then constructed using the LSTM algorithm on the K-means output.

4.1. Feature selection using PCA

PCA is a statistical approach for converting features in a dataset into a new set of uncorrelated features, known as principal components (PCs). The dimensions of a dataset can be reduced using PCA while maintaining the variability of the dataset as much as possible [19]. Several researchers have used PCA with ML techniques. The authors of [20] examined the application of PCA to reduce the dimension of high-dimensional spectral data and improve the performance of various widely used ML models. Their results showed that using the PCA method enhances the performance of ML algorithms for classifying huge amount of data. The researchers of [21] predicted diabetes in patients using three ML techniques: PCA, KMC, and logistic regression (LR). They discovered that, when PCA was used, the KMC algorithm and LR classifiers were more accurate compared to results from previous studies that did not use PCA.

This step seeks to assist in the identification of the most important features that have a significant impact on COVID-19 positive case prediction.

4.1.1. Test for collinearity

The occurrence of a linear correlation between predictors is defined as collinearity. To identify highly collinear predictors, the variance inflation factor (VIF) is employed to assess the strength of the relationship between variables. An increase in the variance of a regression coefficient is indicated using the VIF (Equation (1)) because of collinearity. There is no collinearity when the VIF is less than or equal to one.

$$VIF = 1 / (1 - R\hat{2}) = 1/Tolerance \quad (1)$$

4.2. K-means clustering (KMC)

Clustering is a data-partitioning technique that divides large datasets into smaller groups. Using a distance measure, objects are classified into clusters based on similar features. In this study, K-means clustering was used to identify the outliers. Clustering was accomplished using the following steps:

Step 1. Initialize k = 6. By assigning every input data point to the nearest mean and measuring the similarity using Euclidean distance, we formed six clusters.

Step 2. Centroid/mean value for each cluster's input data is recalculated. Steps (1) and (2) are repeated until the average cluster value converges.

Step 3. Outliers are eliminated by deleting improperly clustered data. This procedure generates a new dataset of variable size s. If s is greater than 70%, then we enter the next phase of classification; otherwise, the KMC procedure is iterated until we obtain an appropriate data size. At the completion of the clustering process, approximately 0.013% of outliers were detected and eliminated.

4.3. LSTM

LSTM is a form of RNN that was expressly built to represent temporal sequences and their long-range interactions better than standard RNNs [22]. Hochreiter and Schmidhuber were the first researchers to propose LSTM in 1997 [23]. LSTM is an RNN enhancement method that substitutes a memory cell for the hidden layer. It was created to alleviate the difficulty associated with learning long-range dependencies between data instances located at great distances from one another in standard RNNs [24].

In Fig. 2, the LSTM concept is used to ensure long-range interdependence by incorporating an intermediate store within the memory

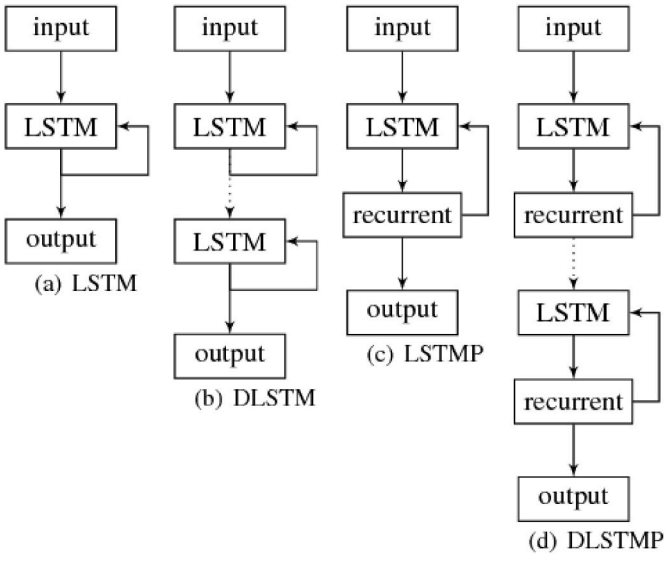

Fig. 2. LSTM Architecture containing memory blocks.

cell, which is controlled by specific neurons called gates. Despite their simplicity, LSTMs perform exceptionally well in various applications. Most LSTMs and their variants are used in RNNs.

Standard LSTM networks include three layers: input, recurrent LSTM, and output. The LSTM layer connects the cell output units to the input units via recurrent connections as well as input, output, and forget gates. The cell output units are connected to the output layer of the network.

In the original architecture, each memory block contains input and output circuits. The input gate controls the flow of the input signals into the memory cell, whereas the output gate controls the flow of the memory cell signals across the network. The forget gate adaptively forgets or resets the cell's memory by scaling its internal state before injecting it as an input via the cell's self-recurrent link [25].

The hidden layer function H in most RNNs is a sigmoid function. H is implemented using the following function in the LSTM version used in this study: Equations (2)–(6) represent the implementation of RNNs.

$$it = \sigma\,(Wxixt + Whiht-1 + Wcict-1 + bi) \tag{2}$$

$$ft = \sigma\,(Wxfxt + Whfht-1 + Wcf\ ct-1 + bf\ ) \tag{3}$$

$$ct = ftct-1 + it\ \tanh(Wxcxt + Whcht-1 + bc) \tag{4}$$

$$ot = \sigma\,(Wxoxt + Whoht-1 + Wcoct + bo) \tag{5}$$

$$ht = ot\ \tanh(ct) \tag{6}$$

where σ represents the logistic sigmoid function, I is the gate, f is the forget gate, and o is the output gate. Subscripts for the matrix of weights are Whi, which denotes the hidden input gate matrix, and Wxo, which denotes the input-output gate matrix. Owing to the diagonal nature of the weight matrices connecting the cell to the gate vectors (e.g., Wci), element m in each gate vector receives the input entirely from the cell vector's element m.

The proposed approach is developed by employing LSTM with an output from PCA and K-means clustering.

4.4. Performance metrics

COVID-19 prediction tool's performance was evaluated using the following performance metrics.

4.4.1. Classification accuracy

This measure expresses the proportion of accurate predictions relative to the total number of input data samples, as expressed in Equation (7).

$$Accuracy = number\ of\ accurate\ predictions \quad / total\ number\ of\ predictions \tag{7}$$

4.4.2. AUC

The AUC is used to evaluate the binary classification. The AUC value effectively summarizes the performance of the receiver operator curves. Researchers of [26] proved that the AUC performance metric is better than the accuracy. Equation (8) represents the formula for sensitivity.

$$Sensitivity = True\ positive\ /\ \_(TrueNegative + FalsePositive) \tag{8}$$

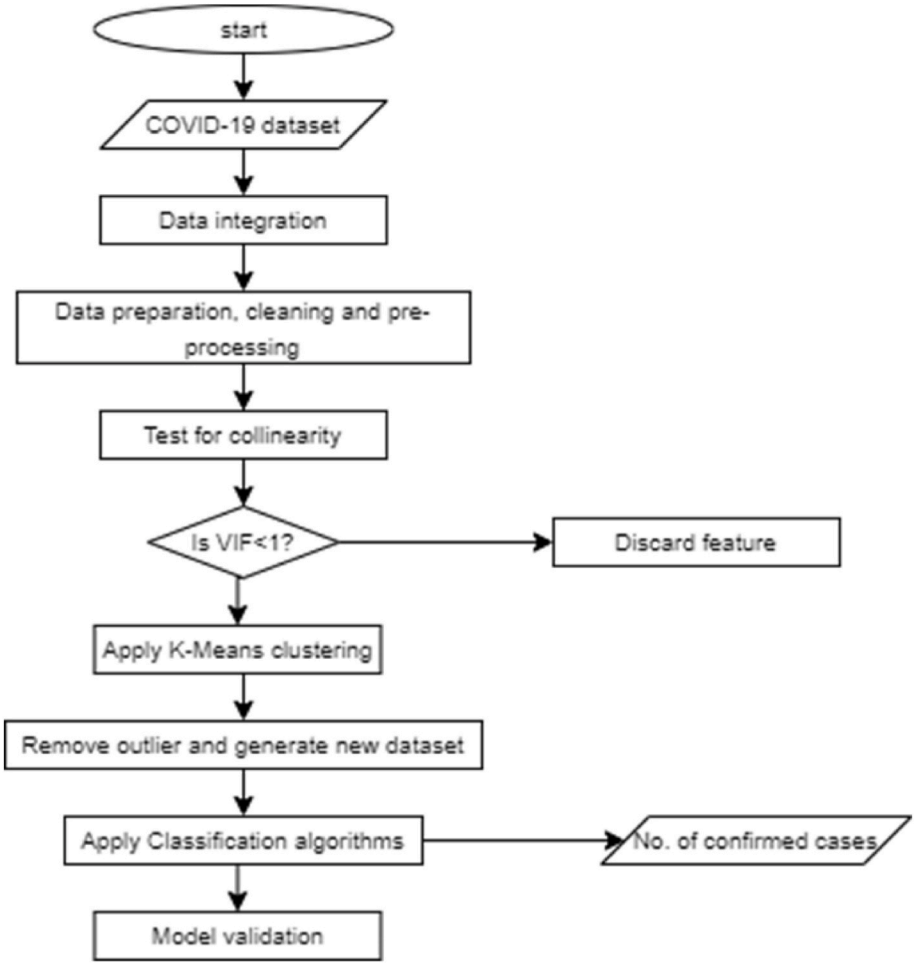

Fig. 3. Flowchart of the COVID-19 prediction model.

$$Specificity = True\ Negative/\ (True\ Negative + False\ Positive) \quad (9)$$

4.5. System implementation

Anaconda 3, which supports Python 3.6 programming language, was used to implement the proposed model. It is an open-source software project consisting of several packages that aid in the development of ML and data science applications. The following steps comprise the design flow of the COVID-19 prediction tool. The flowchart in Fig. 3 is also used to describe the steps used in the implementation of the proposed system.

Step 1. Data preparation and cleaning

i. Procure dataset
ii. Integrate dataset
iii. For each record of the dataset

If $record_{ij}$ is NULL.
Discard record.

Step 2. Principal component analysis

i. Obtain the dataset: Divide the dataset in half, x and y, with x and y representing the validation and training sets, respectively.
ii. Provide a structure to the dataset: Consider the two-dimensional independent variable matrix y, in which the rows and columns correspond to data items and characteristics, respectively. Subtract the column's mean from each entry
iii. Ensure consistency in the data:

$$z = (\boldsymbol{value} - \boldsymbol{mean})/\boldsymbol{standard\ deviation}$$

iv. A covariance matrix is a p-by-p symmetric matrix (where p is the dimension) that contains the covariances of all possible pairings of the initial variables.

Transpose the matrix Z and multiply the resulting transposed matrix by Z.

$$Convariance\ of\ Z = Z^{T}Z$$

The resulting matrix is Z's covariance matrix up to a certain value.

Step 3. Random forest regressor (bagging and boosting)

Step 4. K-means clustering
Apply Section 4.2 to obtain the relevant cluster.

Step 5. Classification and K-fold cross-validation

5. Results and discussion

The experimental results are described in the following section. This section compares the proposed approach with various ML algorithms.

The NCDC COVID-19 dataset was used to predict COVID-19 positive cases in Nigeria, and the suggested model was compared with three ML models: SVM, naive Bayes, and XGBoost. Table 3 shows the results of the comparison of LSTM with other algorithms.

A notable advantage of PCA is that it helps mitigate the problem of redundant characteristics that are useless for clustering. PCA improved our K-means results because it aided in the management of noisy and outlier data by reducing the number of variables in the original dataset. The primary advantage of PCA is that, once we have identified and compressed these principal components from the data (i.e., reducing the number of dimensions without significantly sacrificing information), it becomes necessary to determine the number of clusters and provide a statistical framework for modelling the cluster structure.

First, we employed a K-fold cross-validation technique to assess our model's performance when confronted with new and previously unlearned data. Based on our decision to perform threefold cross-validation, our dataset was divided into three subgroups. In each trial, one subset was used as the test set, and the remaining nine were used as the training set. Th' overall performance of our model was determined by calculating the average error across all 10 trials. This technique resolves two issues: first, it alleviates the bias problem by fitting virtually all the data, and second, it considerably reduces the variance problem.

Table 3
Comparison of ML algorithms.

Algorithm	Accuracy	Sensitivity	Specificity
Naïve Bayes	69%	75%	70%
SVM	92%	89%	90%
LSTM	98.1%	98%	98%
XGBoost	91%	76%	80%

5.1. Comparison of the results

To evaluate the performance of the COVID-19 prediction model, Table 3 compares its average accuracy score with that of various ML classification models, such as naive Bayes, SVM, and XGBoost, on the same dataset with different versions.

The program was executed 20 times and the average accuracy was recorded to avoid any bias in the output.

As presented in Table 3, the approaches of PCA and K-means integration significantly enhanced the performance accuracy of various algorithms employed to model our dataset; LSTM had the greatest accuracy of 98.1%. This may be because the COVID-19 dataset is nonlinear and the LSTM algorithm performs best on a nonlinear dataset. Naïve Bayes and SVM display the same accuracy.

6. Conclusion

By integrating PCA and K-means clustering methods, this paper describes a novel ML-based prediction system capable of identifying positive cases using the NCDC COVID-19 dataset. The results indicate that, compared to other ML methods, LSTM with PCA, K-means, and LSTM has the highest accuracy.

To ensure that the COVID-19 prediction model had a high degree of precision, irrelevant features were removed from the dataset using a feature engineering process, the dataset was cleaned, and outliers were removed through K-means clustering.

The findings of this research will aid in the decision-making regarding proper planning for future COVID-19 outbreaks. Additionally, this approach would assist the governments of Nigeria and other nations in determining the rate of illness spread, thus leading to the development of preventative measures to halt the spread. It can also help with budgeting, particularly when eradication techniques, such as awareness campaigns and the distribution of COVID-19 vaccinations to the public.

Funding

None.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Authors wish to acknowledge African Development Bank and African University of Science & Technology, Nigeria for their support to carry out this research.

References

[1] Iboi EA, Sharomi O, Ngonghala CN, Gumel AB. Mathematical modeling and analysis of COVID-19 pandemic in Nigeria. Math Biosci Eng 2020;17(6):7192–220. https://doi.org/10.3934/MBE.2020369.

[2] Singhal A, Singh P, Lall B, Joshi SD. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos, Solit Fractals 2020;138:110023. https://doi.org/10.1016/j.chaos.2020.110023.

[3] Pai C, Bhaskar A, Rawoot V. Investigating the dynamics of COVID-19 pandemic in India under lockdown. Chaos, Solit Fractals 2020;138:109988. https://doi.org/10.1016/j.chaos.2020.109988.

[4] Rafiq D, Suhail SA, Bazaz MA. Evaluation and prediction of COVID-19 in India: a case study of worst hit states. Chaos, Solit Fractals 2020;139:110014. https://doi.org/10.1016/j.chaos.2020.110014.

[5] Okuonghae D, Omame A. Analysis of a mathematical model for COVID-19 population dynamics in Lagos, Nigeria. Chaos, Solit Fractals 2020;139:110032. https://doi.org/10.1016/j.chaos.2020.110032.

[6] Yadav M, Perumal M, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solit Fractals 2020;139:110050. https://doi.org/10.1016/j.chaos.2020.110050.

[7] Li WT, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inf Decis Making 2020;20(1):247. https://doi.org/10.1186/s12911-020-01266-z.

[8] Bayat V, et al. A severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) prediction model from standard laboratory tests. Clin Infect Dis 2021;73(9): E2901–7. https://doi.org/10.1093/cid/ciaa1175.

[9] Wu J, et al. Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv 2020. https://doi.org/10.1101/2020.04.02.20051136.

[10] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst 2020;44(8). https://doi.org/10.1007/s10916-020-01597-4.

[11] Tschoellitsch T, Dünser M, Böck C, Schwarzbauer K, Meier J. Machine learning prediction of SARS-CoV-2 polymerase chain reaction results with routine blood tests. Lab Med 2021;52(2):146–9. https://doi.org/10.1093/labmed/lmaa111.

[12] Joshi RP, et al. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. J Clin Virol 2020;129 (June):104502. https://doi.org/10.1016/j.jcv.2020.104502.

[13] Shoer S, et al. A prediction model to prioritize individuals for a SARS-CoV-2 test built from national symptom surveys. Med 2021;2(2):196–208. https://doi.org/10.1016/j.medj.2020.10.002. e4.

[14] Tordjman M, et al. Pre-test probability for SARS-Cov-2-related infection score: the PARIS score. PLoS One 2020;15(12 December):1–14. https://doi.org/10.1371/journal.pone.0243342.

[15] Arora P, Kumar H, Panigrahi BK. Prediction and analysis of COVID-19 positive cases using deep learning models: a descriptive case study of India. Chaos, Solit Fractals 2020;139. https://doi.org/10.1016/j.chaos.2020.110017.

[16] Kotsiantis SB, Kanellopoulos D. "Data preprocessing for supervised leaning. Int J … 2006;1(2):1–7. https://doi.org/10.1080/02331931003692557.

[17] Seyed S, Mohammad G, Kamran S. Combination of feature selection and optimized fuzzy apriori rules: the case of credit scoring. Int Arab J Inf Technol 2015;12(2).

[18] Santhanam T, Padmavathi MS. Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci 2015;47(C):76–83. https://doi.org/10.1016/j.procs.2015.03.185.

[19] Jollife IT, Cadima J. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 2016;374(2065). https://doi.org/10.1098/rsta.2015.0202.

[20] Howley T, Madden MG, Connell MO, Ryder AG. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. Appl. Innov. Intell. Syst. 2006;XIII(August 2014). https://doi.org/10.1007/1-84628-224-1.

[21] Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inform Med Unlocked 2019;17(April):100179. https://doi.org/10.1016/j.imu.2019.100179.

[22] Lee D, et al. Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus. China Commun 2017;14(9):23–31. https://doi.org/10.1109/CC.2017.8068761.

[23] Zhang Q, Wang H, Dong J, Zhong G, Sun X. Prediction of sea surface temperature using long short-term memory. Geosci Rem Sens Lett IEEE 2017;14(10):1745–9. https://doi.org/10.1109/LGRS.2017.2733548.

[24] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. 30th Int Conf Mach Learn ICML 2013;(PART 3):2347–55. 2013.

[25] HYBRID SPEECH RECOGNITION WITH DEEP BIDIRECTIONAL LSTM Alex Graves. Navdeep Jaitly and Abdel-rahman Mohamed University of Toronto department of computer science. Canada; 2013. p. 273–8. 6 King ' s College Rd . Toronto , M5S 3G4.

[26] Ling CX, Huang J, Zhang H. AUC: a better measure than accuracy in comparing learning algorithms. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 2003;2671:329–41. https://doi.org/10.1007/3-540-44886-1_25.