Journal of Water & Health

IWA PUBLISHING

Journal of Water and Health Vol 22 No 1, 21 doi: 10.2166/wh.2023.026

How can machine learning predict cholera: insights from experiments and design science for action research

Hauwa Ahmad Amshi [a,*], Rajesh Prasad [b], Birendra Kumar Sharma[c], Saratu Ilu Yusuf[d] and Zaharaddeen Sani[a]

[a] African University of Science and Technology, Abuja, Nigeria
[b] Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India
[c] Ajay Kumar Garg Engineering College, Ghaziabad, India
[d] Bayero University, Kano, Nigeria
*Corresponding author. E-mail: ahauwa@aust.edu.ng

HAA, 0000-0002-0705-4939; RP, 0000-0002-3456-6980

ABSTRACT

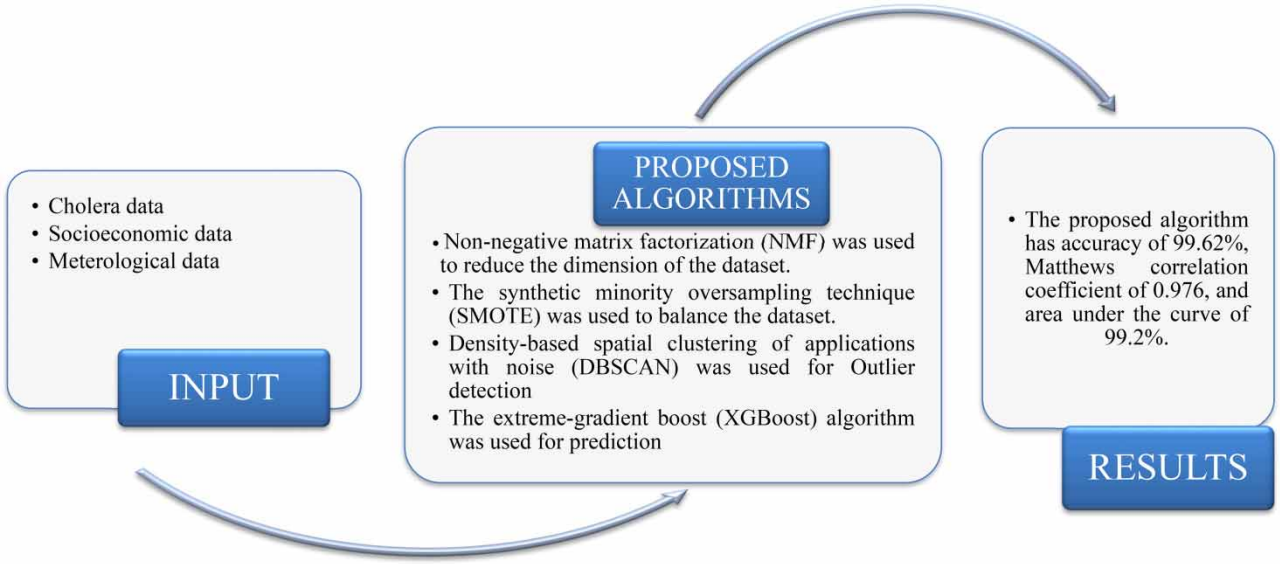
Cholera is a leading cause of mortality in Nigeria. The two most significant predictors of cholera are a lack of access to clean water and poor sanitary conditions. Other factors such as natural disasters, illiteracy, and internal conflicts that drive people to seek sanctuary in refugee camps may contribute to the spread of cholera in Nigeria. The aim of this research is to develop a cholera outbreak risk prediction (CORP) model using machine learning tools and data science. In this study, we developed a CORP model using design science perspectives and machine learning to detect cholera outbreaks in Nigeria. Nonnegative matrix factorization (NMF) was used for dimensionality reduction, and synthetic minority oversampling technique (SMOTE) was used for data balancing. Outliers were detected using density-based spatial clustering of applications with noise (DBSCAN) were removed improving the overall performance of the model, and the extreme-gradient boost algorithm was used for prediction. The findings revealed that the CORP model outcomes resulted in the best accuracy of 99.62%, Matthews's correlation coefficient of 0.976, and area under the curve of 99.2%, which were improved compared with the previous findings. The developed model can be helpful to healthcare providers in predicting possible cholera outbreaks.

Key words: cholera, DBSCAN, dimensionality reduction, NMF, SMOTE and XGBoost

HIGHLIGHTS

- Identifying the cholera prediction attributes.
- Using socioeconomic variables to predict cholera.
- The use of NMFSMOTE for outlier detection and balancing of the dataset.
- Accuracy improved by the use of NMFSMOTE and DBSCAN.
- The developed model can be helpful to healthcare providers in predicting possible cholera outbreaks.

GRAPHICAL ABSTRACT



INTRODUCTION

Cholera is caused by *Vibrio cholerae* 01 or 0139, a bacterium that causes acute watery diarrhea and serious intestinal illness (Asunduwa *et al.* 2020; Deen *et al.* 2020). The exotoxin generated by *Vibrio* can induce dehydration, circulatory failure, shock, and electrolyte imbalance owing to significant (and sometimes fast) fluid and electrolyte loss from the gastrointestinal system. This cycle of events might lead to acidosis, myocarditis, heart failure, tubular necrosis, and death if there is no timely intervention. Cholera is most often disseminated through the intake of contaminated water or food, as a result of poor sanitation, through the fecal-oral pathway (Daisy *et al.* 2020). The majority of cholera infections in developed nations spread via contamination. In undeveloped counties, it is frequently transmitted via polluted water (Miller *et al.* 1985).

The current cholera pandemic, the seventh since 1817, has disproportionately affected Africa, with Nigeria being one of the countries facing significant challenges. Cholera has become prevalent in most African countries over the last 50 years due to various factors such as lack of access to clean water, poor sanitation, natural catastrophes, and internal conflicts that lead to crowded living conditions and interrupting water and sanitation services (Mengel *et al.* 2014; Leckebusch & Abdussalam 2015; Deen *et al.* 2020). As of 31 December 2021, Yobe, Nigeria, reported a peak of 4,003 suspected cholera cases with 91 associated deaths, making it a major cause of concern (Usman & Charangi 2021).

Owing to the lack of emergency or public health response, the detection process of the outbreak might take a longer period, causing the disaster to be more intense. To attain improved health outcomes/disaster management, predictions may be generated by anticipating and assessing risk variables to prevent cholera before it outbreaks. This has prompted the creation of a highly accurate system that examines cholera outbreak data to predict its outbreaks. Technology has been applied to the field of health for monitoring, evaluating, and raising awareness. Social media is being utilized to educate the public, disseminate information, and offer emotional support to those suffering from various conditions such as COVID-19 and the Mpox in Nigeria (Sow *et al.* 2022). The use of social media and mass media for education and awareness is recommended to enhance cholera prevention and control. In a previous study (Willem *et al.* 2014), the authors discussed how modeling plays a major role in policy making, especially for infectious disease interventions. In another previous study (Ibrahim *et al.* 2021), the authors explained how the lack of data prevented individuals, communities, or populations from benefiting from innovations. They discussed how a lack of data poses a threat to global health and might prevent data-driven digital health solutions from being more broadly exploited. Electronic health records include comprehensive data on patient demographics, medical history, current prescriptions, laboratory test results, and diagnosis. Improving patient health information management is the main goal of using electronic health records. Effective machine learning (ML) methods for prediction and forecasting can use electronic health records.

The use of ML to anticipate cholera outbreaks has increased in recent years. ML methods are frequently more accurate and effective than manual classification by subject matter experts (Alfred & Obit 2021). Cholera prediction is a complex problem, but few studies have explored the imbalanced data that impair the performance of classification methods. ML methods have been used in numerous studies to predict cholera using historical data. Medical datasets still present significant challenges owing to their high dimensionality and class imbalance. Using ML without addressing the aforementioned issues decreases the effectiveness of the approaches and, consequently, their accuracy (Fotouhi *et al.* 2019).

The success of ML in other medical, complicated diagnostic, and scientific endeavors has led to its application to identify cholera outbreaks. Therefore, it is anticipated that ML analysis will revolutionize the identification and prevention of cholera outbreaks. Akanda *et al.* (2012) asserted that a reliable and robust cholera prediction model will enable the deployment of expert human (physicians and health workers) and material (vaccine, water purification and sanitation equipment, antibiotics, oral rehydration solution) resources to vulnerable areas in order to prepare for and implement well-planned prevention strategies.

Researchers have developed a risk prediction model for cholera using different models and classifiers in supervised ML. Oversampling might be used to create a balanced dataset owing to limited data (Leo *et al.* 2019). However, further analyses, as well as other preprocessing and feature types using socioeconomic data, need to be conducted (Campbell *et al.* 2020). In addition, when it comes to predicting cholera outbreaks, ML prediction methods outperform conventional methods. The use of ML to identify cholera outbreaks has yielded encouraging results; however, the majority of cholera prediction models still have certain limitations that need to be addressed (Campbell *et al.* 2020). It is necessary to use actual data to validate the outcomes of the categorization prediction models.

ML techniques can predict the risk of cholera outbreaks by automatically identifying a collection of key predictive variables (Asadgol *et al.* 2019). The introduction of ML techniques may help improve the prediction model because current risk assessment models produce unsatisfactory results (Garg & Mago 2021). Prior studies have employed several ML systems to predict cholera outbreaks with a focus on climatic variables with ML models. In a previous study (Chau *et al.*, 2016), daily relative humidity, daily temperature, daily sun hours, daily wind speed, and daily precipitation with random forest (RF) regression were used to obtain a result with a decreased adj-R^2 measure of 0.0076, with the best model being the complete model and the 95% confidence interval being [0.0095, 0.0057].

A previous study (Badkundri *et al.* 2019) proposed the cholera artificial learning model (CALM), which consists of four (Miller *et al.* 1985) extreme-gradient boost (XGBoost) ML prototypes that forecast the incidence rate of new cholera outbreaks in a Yemeni governorate between 2 weeks and 2 months. CALM is a special ML approach that uses data on rainfall, historical cholera incidence, fatality rates, deaths from civil war, and relationships between governorates over a wide time span. However, the aforementioned study had certain limitations in predicting cholera outbreaks using the proposed approaches because of the imbalanced nature of the data.

To overcome the aforementioned limitations, some studies incorporated a balancing technique into their proposed model. Chau (2017)) used the random oversampling examples resampling method on datasets to overcome the disadvantages of an imbalanced dataset and used more seasonal data to increase the prediction performance of the cholera classification mode. Two studies (Leo *et al.* 2019; Campbell *et al.* 2020) used ADASYN and synthetic minority oversampling technique (SMOTE), respectively, to overcome the imbalance problem of the datasets. The overall performance of the model was improved by using resampling methods. The drawback of the aforementioned classification models is that none of the models leverage the use of socioeconomic data, which is a major driver of cholera outbreaks (Çavdaroğlu *et al.* 2022), and clustering techniques for outlier detection and dimensionality reduction were not incorporated in any of these models.

The foundation of the design science paradigm, which is at its core, a paradigm for solving problems, is artificial intelligence and engineering (vom Brocke *et al.* 2020). It seeks to advance human understanding by creating innovative products and design knowledge through innovative solutions to urgent problems (Peffers *et al.* 2007). Because our dataset is highly imbalanced, balancing the data using the oversampling technique will aid in improving the overall performance of the model.

Our aim was to develop a cholera outbreak risk prediction (CORP) modeling ML tools and data science with the following objectives:

i. To determine whether dimensionality reduction using nonnegative matrix factorization (NMF) will improve the performance of the model.
ii. To determine whether balancing the distribution of the training dataset using SMOTE will improve the overall performance of the model.

iii. To determine whether detection and removal of outliers in data using the density-based spatial clustering of applications with noise (DBSCAN) have an effect on the performance of the model.
iv. To develop the prediction model using the ML classification algorithm, XGBoost.
v. To compare with existing state of work models.

The remainder of this article is organized as follows: Section 2 discusses the materials and proposed methods, Section 3 presents the results, Section 4 discusses the experiments and a comparison with the existing literature, and Section 5 discusses the conclusion and future work.

MATERIALS AND METHODS

This section describes the dataset, resampling method, dimensionality reduction method, outlier detection method, and the ML approach used to design this model.

Study site

The study site of this research was Yobe, Nigeria, which is located at an elevation of 347.31 m (1,139.47 ft) above sea level with a latitude of 12.293876, longitude of 11.439041, and area of 45,750 km^2. Yobe is bordered to the east by Borno, to the south by Gombe, to the west by Bauchi and Jigawa, and to the north by Republic of the Niger (Usman & Charangi 2021). Because the state is located in the Sahel-Savanna region, the weather is hot and dry for most of the year, with the exception of the southern part of the state, which receives greater yearly rainfall. The climate is characterized by two distinct seasons: dry and rainy. The rainy season begins in May or June in the south and June or July in the north, peaks in August, and ends in September or October (Kehinde *et al.* 2021). The state has an annual average temperature of 31.29 °C (88.32°F), which is 1.83% higher than Nigeria's average temperature. Usually, Yobe receives about 48.46 mm (1.91 in.) of precipitation with 67.45 rainy days (i.e., 18.48%) annually (Zemba *et al.* 2018). A population of 3,649,600 is projected in 2022, with an annual increase of 2.9% (Nigerian Bureau of Statistics).

Datasets

The dataset used in implementing the CORP model, which includes Yobe cholera data, socioeconomic data, and meteorological data, is discussed in this section.

Cholera dataset

Data used for the implementation of the CORP model were obtained from the Yobe State Ministry of Health. In this study, a secondary source was used for data collection. Therefore, all medical records where cholera outbreaks occurred in the previous years in Yobe were considered. It consists of ID_Number, LGA, ward, settlement, name of patient, age (years), sex (M/F), date of onset, date seen at HF, date lab specimen taken, date sample sent to lab, results for RDT, and outcome 1 = alive, 2 = died, 9 = unknown, 1 = inpatient, 2 = outpatient. Furthermore, a total of 5,855 datasets were collected in this study, with 357 reported cases taking the RDT, making the dataset highly imbalanced. The reported case was then summed up and used as an effective variable type for ML application.

Meteorological data

The meteorological datasets used were obtained from the World Bank Climate Knowledge Portal (https://climateknowledgeportal.worldbank.org/country/nigeria/climate-data-historical) (Climate Knowledge Portal), which consists of historical data, projected data, and general climate variability for different countries. Climatology, timeseries, and heatplot data are provided as a CSV file. The data obtained in a .csv format include the monthly minimum, maximum, and mean temperatures, and precipitation for 2018–2021. Each variable timeseries is downloaded as .csv for annual and monthly records of each variable.

Socioeconomic data

The UNICEF Data Warehouse was used to obtain annual state-level socioeconomic statistics from 2018 to 2021 (https://data.unicef.org/resources/data_explorer/unicef_f/) (UNICEF Data Warehouse). The data acquired include the percentage of people who have access to safe drinking water, number of improved sanitation facilities available, and number of basic hygiene facilities available. Water, Sanitation and Hygiene (WASH) was downloaded as a .csv file which comprises data for national, area, or territory for 2015 and 2020. For each year, the annual rate of change in basic sanitation, water, and

hygiene is specified as 1.10, 0.7, and 0.13%, respectively. The annual rate of change in basic sanitation, access to water, and hygiene is used to estimate the data for 2018, 2019, and 2021.

Proposed framework

The CORP model was implemented using PyCharm, an open-source software that comprises tools for applications in data science and ML that works with Python 3.9. The following steps constitute the design process for the cholera prediction tool, as shown in Figure 1.

Step 1: Preprocess/cleaning and integrating the datasets
Step 2: Apply NMF for dimensionality reduction
Step 3: Apply SMOTE to balance the imbalanced data
Step 4: Use DBSCAN to check for outliers
Step 5: Apply the XGBoost classifier

The proposed model was built and implemented by combining the advantages of the SMOTE, DBSCAN, NMF, and XGBoost models. SMOTE balances the imbalanced data, whereas NMF decreases the dimensionality of these datasets by

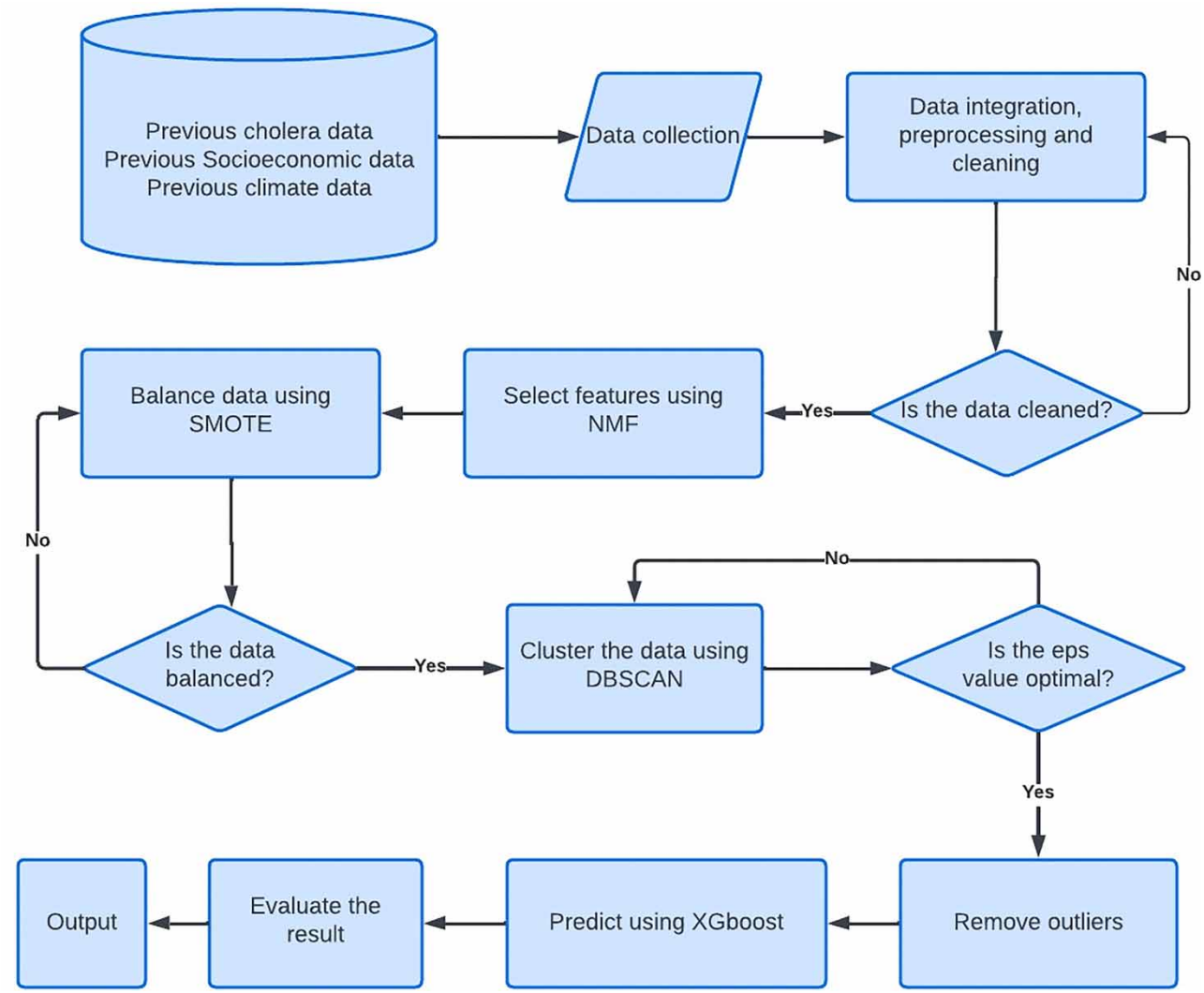


Figure 1 | Flowchart for the design of the CORP model. From a data bank, the datasets are integrated and cleaned, NMF was used for feature selection, and SMOTE was used to address imbalances. DBSCAN identifies potential outliers for enhanced dataset robustness. The final step employs XGBoost, an ensemble classifier, for accurate and efficient classification. This comprehensive pipeline aims to optimize dataset quality, tackle imbalances, and fortify the model against outliers, ensuring a resilient and reliable predictive modeling process.

filtering out any unsuitable data while minimizing information loss, making them easier to analyze (Babaee *et al.* 2016). Unsupervised clustering using DBSCAN was used to remove outliers from the balanced data (Sanguanmak & Hanskunatai 2016). After cleaning the DBSCAN clustering results, we applied the XGBoost technique to build our supervised classification of the cholera dataset.

Data preprocessing

This process involved cleaning, integrating the data obtained from different sources, removing values, or changing features that can negatively impact the model. The purpose of this stage was to transform the data into a form that could be used for model fitting and additional analysis. During the preprocessing step, data from three different sources (i.e., previous cholera data, meteorological data, and socioeconomic data) were checked for errors in data entry, such as missing data. The entire dataset was saved in Microsoft Excel (Microsoft Office 2013 Desktop Publishing Suite; .xls file), which were then integrated into a single dataset. Finally, the information was converted into a comma separated variable (.csv) file. Then, in Python, the features were scaled by a minimum and maximum value (MinMaxScaler) between 0 and 1 to improve the distance-based technique in the dataset.

The following sections explain dimensionality reduction using the NMF method, outlier detection using DBSCAN, and dataset balancing using SMOTE. Figure 2 shows a flowchart of the preprocessing step before proceeding to the dimensionality reduction stage.

Dimensionality reduction using NMF

Owing to the large number of variables in the dataset, dimensionality reduction was required to select the most relevant variables while conserving variance (Kotsiantis & Kanellopoulos 2006). NMF is an unsupervised learning algorithm used to

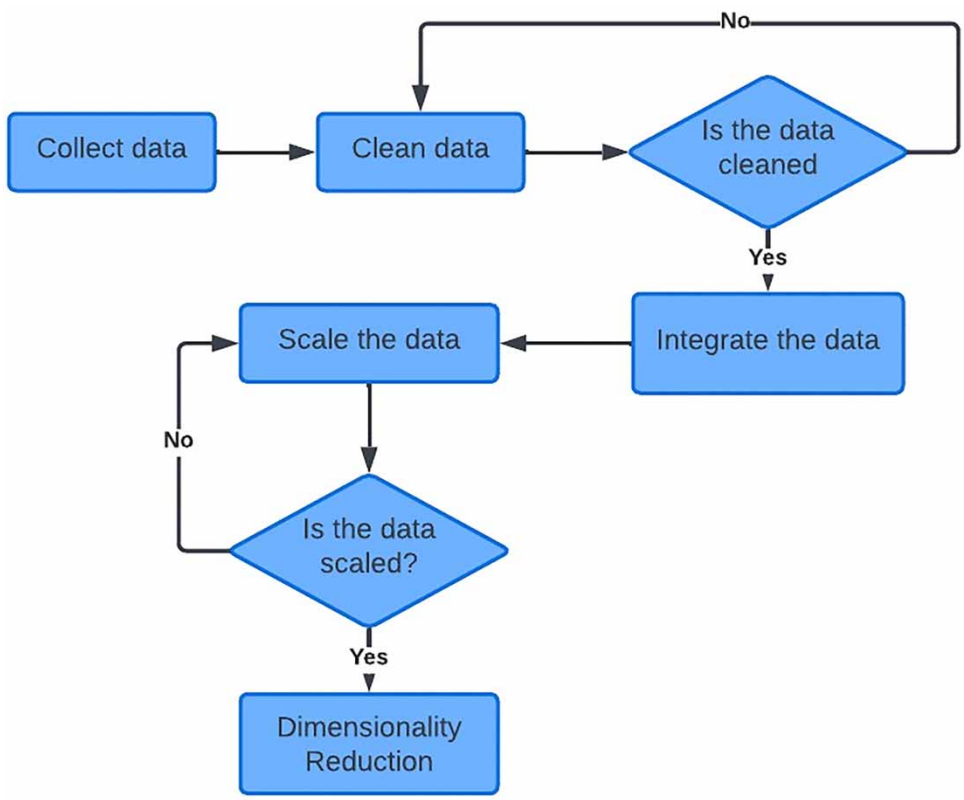


Figure 2 | Data preprocessing in the preprocessing phase, and datasets undergo checks for cleanliness. If already clean, data are directly integrated. Otherwise, it is cleaned before integration. Following integration, the datasets undergo scaling to ensure uniformity. Subsequently, NMF is applied to perform feature selection, streamlining the data for enhanced analysis and model efficiency. This sequential process guarantees data integrity, homogeneity, and optimal feature representation.

reduce data dimensionality to low-dimensional spaces with the goal of dimensionality reduction and feature extraction (Babaee *et al.* 2016; Pathak 2022). A nonnegative matrix is decomposed by NMF into two nonnegative matrices. Dimensionality reduction can be accomplished by projecting the document vectors onto the space produced by these basis vectors, which have fewer dimensions (Zhang *et al.* 2020; Karan *et al.* 2021). The nonnegative restrictions of NMF differentiate it from other approaches, such as principal component analysis and singular value decomposition. These constraints provide an advantage in applications where data are inherently nonnegative and also provide results with better interpretability (Babaee *et al.* 2016). Furthermore, because NMF computation is based on a simple iterative process, it is useful for applications involving large matrices (Buciu 2008). Following is the mathematical definition of NMF and how NMF performs dimensionality reduction. Figure 3 shows the NMF illustration.

Definition of NMF

Let $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{\boldsymbol{K}\times\boldsymbol{N}}$ be such a nonnegative matrix with $\boldsymbol{K} \in \mathbb{N}$ rows and $\boldsymbol{N} \in \mathbb{N}$ columns. The dimensions K and N of the matrix V are usually thought to be large. Given a number $\boldsymbol{R} \in \mathbb{N}$ smaller than both K and N, the goal of NMF is to find two nonnegative matrices $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{\boldsymbol{K}\times\boldsymbol{R}}$ and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{\boldsymbol{R}\times\boldsymbol{N}}$ such that:

$$\mathrm{V} \approx \mathrm{WH} \tag{1}$$

Usually, R known as the rank is chosen to be smaller than K or N, so that W and H are smaller than the original matrix V. This results in a compressed version of the original data matrix, hence reducing it dimensionality.

To find an approximate factorization $\mathrm{V} \approx \mathrm{WH}$, Euclidean distance is used and is defined as follows:

Definition: Given a nonnegative matrix $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{\boldsymbol{K}\times\boldsymbol{N}}$ and a rank parameter $\mathbb{R}$, minimize:

$$\|\boldsymbol{V} - \boldsymbol{WH}\|^2 \tag{2}$$

With respect to $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{\boldsymbol{K}\times\boldsymbol{R}}$ and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{\boldsymbol{R}\times\boldsymbol{N}}$.

It is difficult to find the global minima since $\|\boldsymbol{V} - \boldsymbol{WH}\|^2$ is not convex. Therefore, multiplicative update rule is used to find the local minima.

At each iteration, new H and W are updated given old H and W using the following multiplicative update rules mentioned until convergence. In practice, this means that repeated iteration of the update rules is guaranteed to converge to a locally optimal matrix factorization:

$$\boldsymbol{H}_{[i,j]}^{\boldsymbol{n}+1} \leftarrow \boldsymbol{H}_{[i,j]}^{\boldsymbol{n}} \frac{((\boldsymbol{W}^{\boldsymbol{n}})^{\boldsymbol{T}})\boldsymbol{V}_{[i,j]}}{((\boldsymbol{W}^{\boldsymbol{n}})^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{n}}\boldsymbol{H}^{\boldsymbol{n}})_{[i,j]}} \tag{3}$$

$$\boldsymbol{W}_{[i,j]}^{\boldsymbol{n}+1} \leftarrow \boldsymbol{W}_{[i,j]}^{\boldsymbol{n}} \frac{(\boldsymbol{V}(\boldsymbol{H}^{\boldsymbol{n}+1})^{\boldsymbol{T}})_{[i,j]}}{(\boldsymbol{W}^{\boldsymbol{n}}\boldsymbol{H}^{\boldsymbol{n}+1}(\boldsymbol{H}^{\boldsymbol{n}+1})^{\boldsymbol{T}})_{[i,j]}} \tag{4}$$

Balancing the imbalanced data using SMOTE

Oversampling can be used to generate a balanced dataset when insufficient data are available because an imbalanced data class can affect the prediction accuracy in the majority class (Chawla *et al.* 2002).

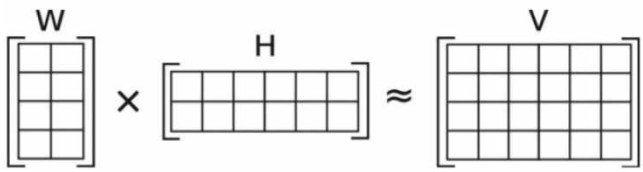


Figure 3 | NMF illustration. The matrix V is represented by the two smaller matrices W and H, which, when multiplied, approximately reconstruct V.

SMOTE is an oversampling strategy widely used in the medical field to address imbalanced datasets (Pavan *et al.* 2020). It is recognized as one of the most dependable and successful preprocessing approach in data analysis and ML (Chawla *et al.* 2002). This is because the new samples are constructed based on real attributes, resembling the real data by creating random synthetic data of the minority class from its nearest neighbors using Euclidean distance, thus increasing data instances (Arafa *et al.* 2022). The number of positive instances (outbreaks) is often significantly smaller than the number of negative instances (nonoutbreaks). This class imbalance can lead to biased model performance and reduced accuracy in predicting cholera outbreaks. Figure 4(a) shows the imbalanced dataset before SMOTE, and Figure 4(b) shows the balanced dataset after SMOTE.

Outlier detection using DBSCAN

Outliers are data points that deviate significantly from the majority of the dataset and can negatively impact model's performance (Reunanen *et al.* 2020). Outliers may arise due to errors in data collection or measurement, and their presence can distort the learning process of the model (Guan & Tibshirani 2022). Removing outliers aims to improve the robustness and accuracy of the model by reducing the influence of extreme values.

DBSCAN is a density-based clustering algorithm that identifies clusters, regardless of the shape and size, existing within a dataset. Point N is the noise point owing to its inaccessibility from any other point based on the fundamental notions of DBSCAN. Point A is the core point, and points B and C are the boundary points because both are densely connected and accessible from point A (Peng & Park 2022). DBSCAN requires only two parameters to generate a new cluster: the radius of the cluster (eps) and the smallest number of points (*MinPts*) inside the eps-radius circle. All points are divided into three groups: core points (x), border points (y), and noise/outlier points (z) (Sanguanmak & Hanskunatai 2016). Figure 5(a) depicts the fundamental concepts underlying DBSCAN, and Figure 5(b) depicts the concept of DBSCAN with $MinPts = 5$ and eps = radius of the circle.

A primary benefit of using DBSCAN as a clustering technique is its ability to identify remote samples that are not accessible from any other location and label them as noise (Arafa *et al.* 2022). Therefore, DBSCAN is considered one of the most

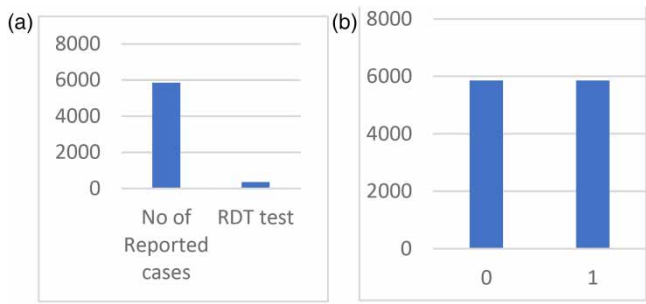


Figure 4 | The class attribute of datasets showing balanced and imbalanced class. A balanced class distribution ensures roughly equal representation of each class, fostering fair model training. Conversely, an imbalanced class distribution indicates a notable discrepancy in class instances, posing challenges for predictive modeling. (a) Imbalanced dataset. (b) Upper-sampled dataset.

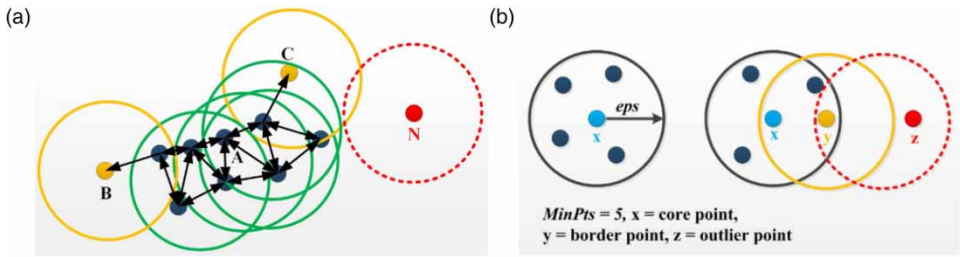


Figure 5 | (a) Fundamental concepts of DBSCAN. (b) Concept of DBSCAN with $MinPts = 5$. The foundational principles of DBSCAN are illustrated, highlighting the concepts of core points, border points, and noise, and showing how adjusting the MinPts parameter influences the clustering behavior.

effective clustering techniques for data mining applications to identify outliers (Sanguanmak & Hanskunatai 2016; Peng & Park 2022). Another benefit of DBSCAN is that it automatically determines the clusters that are present in a given dataset; therefore, the user does not need to determine the number of clusters externally (Latif *et al.* 2020; Arafa *et al.* 2022).

The k-distance graph is used to calculate the optimal eps value by averaging the distances between each point and its k-nearest neighbors (Latif *et al.* 2020) with the user-defined *MinPts* value equivalent to the value of k. In this study, $k = 5$ was used to determine the ideal eps value, as used in previous studies (Zhang *et al.* 2020; Kehinde *et al.* 2021; Pathak 2022), obtaining an eps value of 9 and outlier detection of 6. After removing the detected outliers, we moved to training and model development, which greatly improved the model accuracy for all datasets, increasing the accuracy from 96.71 to 99.62%.

Training and model development

The CORP model was developed using PyCharm, which supports Python 3.8. It is an open-source software package that supports ML and data science applications. The ML classifiers used in this study were RF, Naïve Bayes (NB), and XGBoost. Finally, we used 10-fold cross-validation to choose optimal hyperparameter values. Cross-validation is used because it generalizes ML model's capabilities and is widely used to guarantee that model's performance is dependable and unbiased when applied to fresh, unknown data (Santos *et al.* 2018).

Evaluation Metrics

Accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the curve (AUC) were the metrics used to evaluate the performance of the CORP model.

The accuracy metrics were used to evaluate the model's prediction ability, which took into consideration all components such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FNs). Accuracy was defined as the ratio of the number of accurate predictions to the total number of predictions. Equation (5) provides a mathematical representation of the accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

Sensitivity indicates how well the model performs in terms of the number of positives (cholera outbreaks) correctly identified, as shown in Equation (6):

$$\text{Sensitivity} = \frac{\text{TN}}{\text{TN} + \text{TP}} \tag{6}$$

Complementary to sensitivity is specificity, which determines how well the model performs negatively when there is no outbreak, as shown in Equation (7):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

MCC is a contingency matrix technique for generating the Pearson product moment correlation coefficient between the actual and predicted values that is unaffected by unbalanced datasets. It is mathematically defined in Equation (8) as follows:

$$\text{MCC} = \frac{\text{TP.TN} - \text{FP.FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times}} (\text{TN} + \text{FN}) \tag{8}$$

MCC falls within the range [1, +1], with the extreme values of 1 and +1 obtained in the case of perfect misclassification and perfect classification, respectively, and MCC = 0 being the predicted value for a random guess classifier.

The receiver operating characteristic (ROC) curve was used for the experimental measurements of binary class cases. It is used to display the performance of a binary classifier and shows the trade-off between the TP rate (TPR) and FP rate (FPR). ROC assesses the performance of a classification model. For classifier predictions, TPR versus FPR is displayed. The area under the plot is then computed and called the AUC, which was utilized as an overview of the ROC curve and taken as a

measure of a classifier's ability to distinguish between several classes of plots. An AUC value close to 1 indicates that the model is powerful and has a high level of distinction. An AUC near 0 indicates the worst measure of separability and denotes a poor model. This implies that the results are reversed. Both 1 and 0 s were predicted to be 1. AUC = 1 further indicates that the classifier can properly differentiate between all positive and negative class points.

RESULTS AND DISCUSSION

The precipitation, humidity, minimum and maximum temperatures, and the number of reported cases are independent variables, and the number of cholera cases is the dependent variable; independent variables predict the values of the dependent variable in the model which is the number of cases reported. The primary prediction revolves around identifying potential future occurrences of cholera disease outbreaks. The method used historical data to recognize patterns and trends associated with past outbreaks.

All models were cross-validated using 10-folds, and five performance measures were collected: accuracy, sensitivity, specificity, MCC, and ROC-AUC. MCC provides an effective solution to the class imbalance problem, which is used to evaluate the performance of classification for imbalanced datasets. MCC generates a high score only if the classifier correctly predicts most of the positive and negative data instances, and if most of its positive and negative predictions are correct, then it is considered the best evaluation metric (Reunanen *et al.* 2020). ROC curve was used to visually assess the performance of the classification model. The curve assisted in determining a threshold level that balances the sensitivity and specificity for a given situation. A perfect classifier has an ROC-AUC of 1, whereas a random classifier has an ROC-AUC of 0.5 (Guan & Tibshirani 2022). Figure 6 depicts the ROC curve for the proposed CORP model.

The experiment to verify and evaluate the results of the proposed methods in this section was conducted in two stages. First, the performance of the plain classifiers was compared with that of NMFSMOTE before and after outlier removal. The performance of the trained classifiers on both datasets was compared. Some classes performed well on the evaluation metrics, while others performed poorly. Table 1 compares the outcomes of several classifiers before and after applying NMFSMOTE and also compares the outcomes of NMFSMOTE and DBSCAN, and Table 2 compares the performance of the CORP model before and after outlier detection.

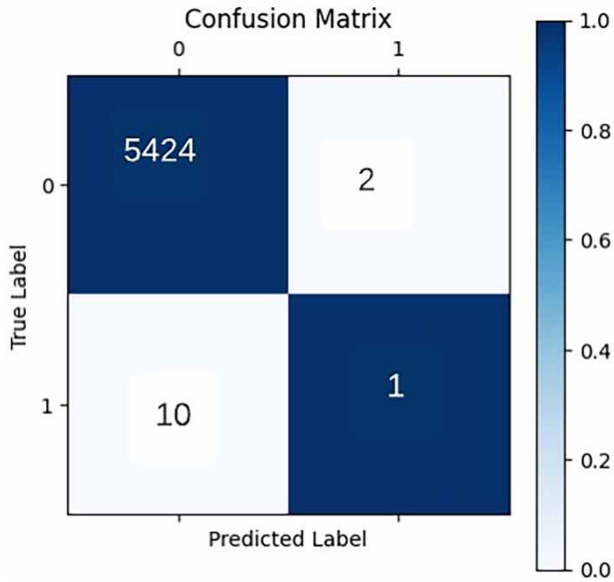


Figure 6 | Confusion matrix: True positive (TP): indicates that the model correctly predicted 5,424 out of 5,855 instances as nonoutbreak. True negative (TN): 423 indicates that the model correctly predicted 423 out of 5,855 instances as outbreak. False positive (FP): 10 indicates that the model incorrectly predicted 10 out of 5,855 instances as nonoutbreak. These are false alarms. False negative (FN): indicates that the model missed one positive case in Yobe State, Nigeria.

Table 1 | Comparison of performance of the selected algorithms before and after balancing

Classifiers		NB	RF	XGBoost
Plain classifiers	Accuracy	90	81.3	81.3
	Sensitivity	0.9	0.72	0.89
	Specificity	0.8	0.6	0.8
	MCC	0.676	0.546	0.8
	AUC	0.63	0.86	0.744
NMFSMOTE	Accuracy	90.2	93.55	96.71
	sensitivity	0.94	0.95	0.91
	specificity	0.9	0.7	0.86
	MCC	0.885	0.92	0.939
	AUC	0.957	0.981	0.99

Table 2 | Comparison of the performance of the CORP model before and after outlier detection

Classifiers		NB	RF	CORP (proposed model)
NMFSMOTE	Accuracy	90.2	93.55	96.71
	Sensitivity	0.94	0.95	0.91
	Specificity	0.9	0.7	0.86
	MCC	0.885	0.86	0.939
	AUC	0.957	0.981	0.99
NMFSMOTE + DBSCAN	Accuracy	96.39	98.2	99.62
	Sensitivity	0.93	0.95	0.93
	Specificity	0.86	0.8	0.89
	MCC	0.80	0.771	0.976
	AUC	0.981	0.997	0.992

The findings revealed that the prediction accuracy increased compared with the other models. For comparison, three ML algorithms, NB, XGBoost, and RF, which have a good record of efficiency and accuracy, have been widely used in previous studies.

The results showed that incorporating NMFSMOTE as a preprocessing step improved the performance of all five ML algorithms. Table 2 shows that the accuracy improved from 90 to 90.2% for NB, 81.3 to 93.55% for RF, and 81.3 to 96.71% for XGBoost. MCC and AUC scores also improved for each of the mode, which indicates that imbalanced datasets hinder the performance accuracy of the model.

The result (Table 2) from XGBoost showed that the mean validation accuracy was 99.62%. This exceptional accuracy signifies the model's proficiency in correctly classifying instances. Notably, the removal of outliers led to a 2.91% improvement in accuracy, emphasizing the significance of outlier detection in enhancing predictive performance. The specificity and sensitivity scores were also improved (0.89 and 0.93, respectively), indicating that 93% of the outbreak was correctly identified. A high sensitivity (0.93) indicates that the model is effective in capturing and correctly classifying instances of cholera outbreaks. It has a strong ability to identify TP cases, which is crucial in the context of disease prediction. A high specificity (0.89) suggests that the model also correctly identify instances where there is no cholera outbreak. It has a strong ability to avoid false-positive predictions, reducing the chances of unnecessary alarms. These enhancements indicate that our model is highly effective in correctly identifying both nonoutbreak and outbreak instances, with a particular emphasis on the latter. The positive impact on sensitivity is crucial for our application, as it ensures that a substantial portion (93%) of actual outbreaks is correctly identified as shown in Figure 6.

There was also a slight improvement of 0.002% with an AUC of 0.992 as shown in Figure 7, indicating that the model is very promising. MCC also improved and had a score of 0.976, indicating the reliability of our classifiers. The MCC further supports the reliability of our classifiers, with considering the balance between TP and TN instances, providing a comprehensive measure of predictive performance. The result from NB showed that the accuracy was 90.2%, with a specificity of 0.86,

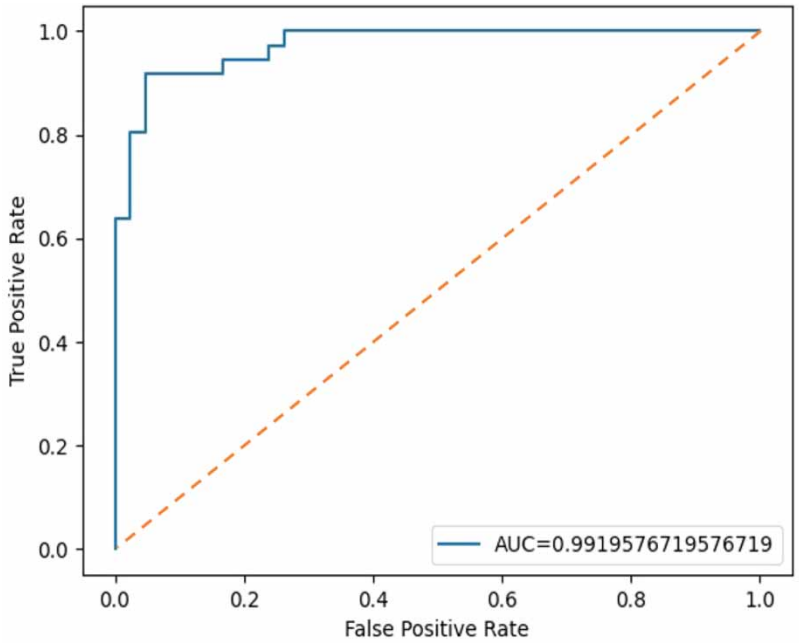


Figure 7 | ROC curve for the proposed CORP model showing an AUC = 0.992 showing a highly effective model, providing a high level of confidence in its ability to correctly classify instances in a binary classification problem.

sensitivity of 0.93, and AUC of 0.981. Furthermore, the RF result showed an accuracy of 98.2%, specificity score of 0.80, sensitivity of 0.95, and AUC of 0.94. After removing the outliers, the performance of all models improved (Table 2).

COMPARISON OF THE RESULTS WITH THE CURRENT MODELS

This section compares the CORP model with already existing cholera prediction models in the literature. Table 3 compares the CORP model existing model.

The findings of the current (benchmark) study (Leo *et al.* 2019) were compared with those of this study. The accuracy of the benchmark study was 99%, with a total of 8,504 reported cases of cholera and an ROC-AUC score of 0.987. Although the benchmark model also used imbalanced datasets, it did not consider the use of MCC as a performance metric. Our model outperformed the benchmark model with some improvements in AUC, accuracy, and sensitivity. Table 3 shows a comparison of the CORP model with existing models.

The XGBoost classifier was chosen as the top model for this study based on the main goal of the study. This is because the XGBoost method is a gradient-boosted decision tree implementation created to be extremely effective, adaptable, and

Table 3 | Comparison of the CORP model with the existing models

Author	Campbell *et al.* (2020)	Chau (2017)	Leo *et al.* (2019)	Proposed (CORP) model
Variables used	Satellite-based climate variables and previous cholera incidence data	Climate variables and previous cholera incidence data	Climate variables and previous cholera incidence data	Climate variables and previous cholera incidence data and socioeconomic data
AUC	0.982	0.84	–	0.992
Accuracy	0.99	–	0.767	0.996
Sensitivity	0.895	0.746	0.805	0.93
Specificity	–	0.778	0.73	0.89
MCC	–	–	–	0.976

portable with the capacity to speed up execution and improve the model's performance. It is particularly useful for anomaly detection in supervised environments where data are frequently imbalanced, such as cybersecurity, DNA sequencing, and credit card transactions.

The XGBoost approach also features a parameter called 'scale-pos-weight,' which focuses on the sensitivity of the data and offers a sequential approach to dealing with imbalanced data. In addition, because it incorporates decision trees into its processes, XGBoost is helpful in decision-making and thus aligns with the study's main objective (Li *et al.* 2020), in contrast to RF, which performs poorly with imbalanced datasets and high-dimensional data. XGBoost outperforms RF in terms of performance, training speed, and accuracy (Bentéjac *et al.* 2021). Furthermore, one of the most notable differences between XGBoost and RF is that, while decreasing model cost, XGBoost always prioritizes functional space, whereas RF tries to prioritize hyperparameters to optimize the model. Based on these succinct descriptions of the two models, the XGBoost classifier was chosen as the best model for this study.

CONCLUSIONS

The goal of this study was to create an ML model that predicts cholera outbreaks, called CORP, to identify cholera outbreaks in Yobe, Nigeria, by combining data science approaches with ML. NMF was used for dimensionality reduction. SMOTE is a data balancing technique. DBSCAN detects outliers, and the XGBoost technique is used for prediction. The results showed that combining the NMFSMOTE, DBSCAN, and XGBoost outcomes resulted in the highest accuracy of 99.62% and AUC of 99.2%, which are superior to those of the current study (98.2 and 99%, respectively). The results presented in Table 2 demonstrate the effectiveness of our approach in predicting outbreaks in Yobe State.

The proposed XGBoost model demonstrates exceptional performance with an accuracy of 99.62%, indicating its ability to reliably distinguish between outbreak and nonoutbreak situations. With a sensitivity of 0.93, the model effectively captures actual outbreaks, minimizing the risk of missing critical events. In addition, the model's specificity of 0.89 ensures that false alarms are raised in only 11% of cases, preventing unnecessary panic and resource allocation. These metrics translate into concrete predictions, enabling the model to pinpoint Yobe State as a high-risk zone.

The findings of this research will contribute significantly to the theoretical growth of the cholera prediction literature. This research will be useful to public health practitioners in anticipation of the potential risk of cholera outbreak, thus making a move ahead. The 'Global Task Force on Cholera Control' (GTFCC) of the World Health Organization (WHO) is supported by this knowledge-based approach, as well as the aim of eradicating cholera by 2030. As a quick method for CORP knowledge discovery that can accommodate a large number of electronic health records and support clinical decision-making, this study has shown the value of electronic health records employing data mining methods. The CORP model would need minimum equipment and operational costs while allowing for the early application of preventative measures to limit disease spread.

To further this work, health systems, particularly in Africa, should adopt appropriate big data and cloud computing practices that improve data capturing, storing, analyzing, sharing, and managing patient data. This will help prediction/diagnosis applications scale more effectively for better health data analysis. More data from longitudinal surveys can be used in future investigations. Adopting deep learning models, dimensionality techniques and other hyperparameter tuning will expand the scope of this study.

And also, to further demonstrate the effectiveness of the CORP model in predicting current and future cholera outbreaks, we will be conducting validation on a separate dataset. By evaluating the model's performance on unseen data, we can assess its generalization capabilities and ensure that it remains reliable and unbiased when applied to real-world scenarios. The enhanced accuracy and reliability observed in our results pave the way for future work in real-time outbreak prediction, as the methodology has demonstrated its potential in identifying and classifying outbreak instances with high precision.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Akanda, A. S., Jutla, A. S., Gute, D. M., Evans, T. & Islam, S. 2012 Reinforcing cholera intervention through prediction-aided prevention. *Bull. World Health Organ.* **90** (3), 243–244.

Alfred, R. & Obit, J. H. 2021 *The Roles of Machine Learning Methods in Limiting the Spread of Deadly Diseases: A Systematic Review*, Vol. 7. Elsevier Ltd, Heliyon.

Arafa, A., El-Fishawy, N., Badawy, M. & Radad, M. 2022 RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *J. King Saud Univ. – Comput. Inf. Sci.* **34** (8), 5059–5074. https://doi.org/10.1016/j.jksuci.2022.06.005.

Asadgol, Z., Mohammadi, H., Kermani, M., Badirzadeh, A. & Gholami, M. 2019 The effect of climate change on cholera disease: The road ahead using artificial neural network. *PLoS One* **14** (11), e0224813. https://doi.org/10.1371/journal.pone.0224813.

Asunduwa, K., Usman, A., Isyaku, A., Shehu, A., Francis, O., Balogun, M. & Aworh, M. K. 2020 Descriptive analysis of a cholera outbreak in 14 LGAs of Sokoto State – Nigeria, 2018. *Int. J. Infect. Dis.* **101**, 363. https://doi.org/10.1016/j.ijid.2020.09.951.

Babaee, M., Tsoukalas, S., Babaee, M., Rigoll, G. & Datcu, M. 2016 Discriminative nonnegative matrix factorization for dimensionality reduction. *Neurocomputing* **173**, 212–223. http://dx.doi.org/10.1016/j.neucom.2014.12.124.

Badkundri, R., Valbuena, V., Pinnamareddy, S., Cantrell, B. & Standeven, J. 2019 Forecasting the 2017-2018 Yemen Cholera Outbreak with Machine Learning, pp. 1–27. Available from: http://arxiv.org/abs/1902.06739.

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. 2021 A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

Buciu, I. 2008 Non-negative matrix factorization, a new tool for feature extraction: Theory and applications. *Int. J. Comput. Commun. Control.* **3** (SPL. ISS.), 67–74.

Campbell, A. M., Racault, M. F., Goult, S. & Laurenson, A. 2020 Cholera risk: A machine learning approach applied to essential climate variables. *Int. J. Environ. Res. Public Health.* **17** (24), 1–24.

Çavdaroğlu, S., Aktar, I., Hasan, M. M., Costa, A. C. dos S., Aborode, A. T., Ahmad, S. & Essar, M. Y. 2022. Cholera amidst COVID-19 pandemic: African healthcare system in jeopardy. Einstein (*São Paulo*) **20**. https://doi.org/10.31744/einstein_journal/2022CE6938.

Chau, N. H. & Ngoc Anh, L. T. 2016 Using Local Weather and Geographical Information to Predict Cholera Outbreaks in Hanoi, Vietnam. In: Nguyen, T.B., van Do, T., An Le Thi, H., Nguyen, N.T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 453. Springer, Cham. https://doi.org/10.1007/978-3-319-38884-7_15.

Chau, N. H. 2017 Enhancing cholera outbreaks prediction performance in Hanoi, Vietnam using solar terms and resampling data, In: Nguyen, N., Papadopoulos, G., Jędrzejowicz, P., Trawiński, B., Vossen, G. (eds) Computational Collective Intelligence. ICCCI 2017. Lecture Notes in Computer Science(), vol 10448. Springer, Cham. https://doi.org/10.1007/978-3-319-67074-4_26.

Chawla, N. V., Bowyer, K. W. & Hall, L. O. 2002 SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intel. Res*, **16** (16), 321–357. https://doi.org/10.1613/jair.953.

Climate Knowledge Portal. Available from: https://climateknowledgeportal.worldbank.org/.

Daisy, S. S., Saiful Islam, A. K. M., Akanda, A. S., Faruque, A. S. G., Amin, N. & Jensen, P. K. M. 2020 Developing a forecasting model for cholera incidence in Dhaka megacity through time series climate data. *J. Water Health.* **18** (2), 207–223.

Deen, J., Mengel, M. A. & Clemens, J. D. 2020 Epidemiology of cholera. *Vaccine*. **38**, A31–A40. https://doi.org/10.1016/j.vaccine.2019.07.078.

Fotouhi, S., Asadi, S. & Kattan, M. W. 2019 A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **90**, 103089. https://doi.org/10.1016/j.jbi.2018.12.003.

Garg, A. & Mago, V. 2021 Role of machine learning in medical research: A survey. *Comput. Sci. Rev.* **40**, 100370. https://doi.org/10.1016/j.cosrev.2021.100370.

Guan, L. & Tibshirani, R. 2022 Prediction and outlier detection in classification problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **84** (2), 524–546. https://doi.org/10.1111/rssb.12443.

Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D. & Denniston, A. K. 2021 Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health* **3** (4), e260–e265. https://doi.org/10.1016/S2589-7500(20)30317-4.

Karan, B., Sahu, S. S., Orozco-Arroyave, J. R. & Mahto, K. 2021 Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction. *Comput. Speech Lang.* **69**, 101216. https://doi.org/10.1016/j.csl.2021.101216.

Kehinde, O. M., Lawan, B. & Umar, A. A. 2021 Changing patterns of temperature in Yobe State, North-Eastern Nigeria: An evidence of climate change. *Asian J. Geographic. Res.* **4** (1), 52–72. https://doi.org/10.9734/AJGR/2021/v4i130126.

Kotsiantis, S. B. & Kanellopoulos, D. 2006 Data preprocessing for supervised leaning. *Int. J. Comput. Sci.* **1** (2), 1–7. Available from: http://www.google.com/search?client=safari&rls=en&q=Data+Preprocessing+for+Supervised+Leaning&ie=UTF-8&oe=UTF-8%5Cnpapers2://publication/uuid/AA4424CD-8BE0-43AB-838B-8BBBDE502355.

Latif, N., Syafrudin, M., Alfian, G. & Rhee, J. 2020 HDPM: An effective heart disease prediction model for a clinical decision support system. IEEE Access, 8, 133034–133050. https://doi.org/10.1109/ACCESS.2020.3010511.

Leckebusch, G. C. & Abdussalam, A. F. 2015 Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Heal Place.* **34**, 107–117.

Leo, J., Luhanga, E. & Michael, K. 2019 Machine learning model for imbalanced cholera dataset in Tanzania. *Sci. World J.* **2019**, Article ID 939757. https://doi.org/10.1155/2019/9397578.

Li, W. T., Ma, J., Shende, N., Castaneda, G., Chakladar, J., Tsai, J. C., Apostol, L., Honda, C. O., Xu, J., Wong, L. M., Zhang, T., Lee, A., Gnanasekar, A., Honda, T. K., Kuo, S. Z., Yu, M. A., Chang, E. Y., Rajasekaran, M. R. & Ongkeko, W. M. 2020 Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. *BMC Med. Inform. Decis. Mak.* **20** (1), 247. https://doi.org/10.1186/s12911-020-01266-z.

Mengel, M. A., Delrieu, I., Heyerdahl, L. & Gessner, B. D. 2014 Cholera outbreaks in Africa. In: Nair, G., Takeda, Y. (eds) *Cholera Outbreaks. Current Topics in Microbiology and Immunology*, vol 379. Springer, Berlin, Heidelberg. 117–144. https://doi.org/10.1007/82_2014_369.

Miller, C., Feachem, R. & Drasar, B. 1985 Cholera epidemiology in developed and developing countries: New thoughts on transmission, seasonality, and control. *Lancet* **325** (8423), 261–263. Available from: https://www.sciencedirect.com/science/article/pii/S0140673685910360.

Nigerian Bureau of Statistics. Available from: https://www.nigerianstat.gov.ng/.

Pathak, S. 2022 Non-negative matrix factorization framework for dimensionality reduction and unsupervised clustering. *Insight J.* **03006**, 1–5.

Pavan, V., Turlapati, K. & Ranjan, M. 2020 Intelligence-based medicine outlier-SMOTE: A re fi ned oversampling technique for improved detection of COVID-19. *Intell. Med.* **3–4**, 100023. https://doi.org/10.1016/j.ibmed.2020.100023.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. 2007 A design science research methodology for information systems research. *J. Manag. Inf. Syst.* **24** (3), 45–77.

Peng, C. Y. & Park, Y. J. 2022 A new hybrid under-sampling approach to imbalanced classification problems. *Appl. Artif. Intell.* **36** (1). https://doi.org/10.1080/08839514.2021.1975393.

Reunanen, N., Räty, T., Jokinen, J. J., Hoyt, T. & Culler, D. 2020 Unsupervised online detection and prediction of outliers in streams of sensor data. *Int. J. Data. Sci. Anal.* **9** (3), 285–314. https://doi.org/10.1007/s41060-019-00191-3.

Sanguanmak, Y. & Hanskunatai, A. 2016 DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification. *13th International Joint Conference on Computer Science and Software Engineering*, JCSSE. Khon Kaen, Thailand, 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748928.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. & Santos, J. 2018 Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. *IEEE Comput. Intell. Mag.* **13** (4), 59–76.

Sow, A., Haruna, U., Abimbola, A., Olajide, E., Amene, T., Odususi, O., Adewusi, B. A., Abia, C., Safari, J., Sorinola, F. W., Alaka, H. O. & Musa, S. M. 2022 Tackling cholera outbreak amidst COVID-19 pandemic in Nigeria: Challenges and recommendations. *Public Health Rev.* **43**, 1604776.

UNICEF Data Warehouse. Available from: https://data.unicef.org/resources/data_explorer/unicef_f/.

Usman, H. & Charangi, F. U. 2021 Spatial distribution of cholera cases in Damaturu Town, Yobe State, Nigeria. *Int. J. Adv. Res. Ideas Innov. Technol.* **7** (5), 150–161.

vom Brocke, J., Hevner, A. & Maedche, A. 2020 Introduction to design science research. Progress in IS. In: Jan vom Brocke, J., Alan Hevner, A. & Maedche, A. (eds.), Design Science Research. Cases, pages 1–13, Springer.

Willem, L., Stijven, S., Vladislavleva, E., Broeckhove, J., Beutels, P. & Hens, N. 2014 Active learning to understand infectious disease models and improve policy making. *PLoS Comput. Biol.* **10** (4), 1–10.

Zemba, A. A., Umar, Y. & Binbol, N. L. 2018 Climatic information as evidence of desertification processes in northern Yobe state, Nigeria: Implications for agriculture and ecosystem. *Global J. Pure Appl. Sci.* **24** (1), 117. https://doi.org/10.4314/gjpas.v24i1.14.

Zhang, S., Yang, L., Yang, J., Lin, Z. & Ng, M. K. 2020 Dimensionality reduction for single cell RNA sequencing data using constrained robust non-negative matrix factorization. *NAR Genomics Bioinforma.* **2** (3), 1–11.

First received 31 January 2023; accepted in revised form 10 December 2023. Available online 20 December 2023