2026 Volume: 02 Issue: 03 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-19

MACHINE LEARNING–BASED PRECISION DIAGNOSIS OF *VIBRIO CHOLERAE* O1 AND O139 STRAINS FOR PERSONALIZED CHOLERA MANAGEMENT

***Okorie Kingsley Maduabuchi, Oko Gabriel Ota**

Computer Science Department, Enugu State University of Science and Technology.

Article Received: 8 February 2026
Article Revised: 28 February 2026
Published on: 20 March 2026

***Corresponding Author: Okorie Kingsley Maduabuchi**
Computer Science Department, Enugu State University of Science and Technology.

DOI: https://doi-doi.org/101555/ijrpa.1787

ABSTRACT

Cholera remains a persistent public health challenge in Nigeria, driven by recurrent outbreaks, poor sanitation, environmental factors, and the evolving genetic diversity of *Vibrio cholerae*, particularly the toxigenic O1 and O139 serogroups. Conventional diagnostic techniques such as culture, polymerase chain reaction, and serotyping, although reliable, are time-consuming, resource-intensive, and inadequate for rapid strain-level differentiation required for effective outbreak response and personalized clinical management. This study proposes a machine learning–based precision diagnostic framework for the early detection and strain-specific classification of *Vibrio cholerae* O1 and O139 infections, with a focus on the Nigerian context. Clinical and environmental data obtained from confirmed cholera cases provided by the Nigeria Centre for Disease Control (NCDC), Lagos State branch, were preprocessed through data cleaning, feature selection, normalization, and class balancing techniques. A Random Forest (RF) classifier was developed and evaluated alongside a K-Nearest Neighbors (KNN) model using a 70:30 training–testing split. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. The Random Forest model demonstrated superior performance, achieving an accuracy of 85.2%, precision of 0.867, recall of 0.867, and F1-score of 0.867, outperforming the KNN model across all evaluation metrics. The findings underscore the potential of machine learning–driven precision diagnostics in enhancing cholera detection, strain differentiation, and risk prediction in resource-limited settings. This study contributes to advancing the application of artificial intelligence in infectious disease genomics and supports the integration of data-driven diagnostic tools into cholera surveillance, clinical decision-making, and public health preparedness strategies in Nigeria.

1. INTRODUCTION

***1.1 Overview*.** Cholera is a severe diarrhoeal disease caused by toxigenic variants of *Vibrio cholerae*, primarily the O1 and O139 serogroups (Tarh, 2020). The O1 serogroup—especially the seventh pandemic El Tor lineage—continues to drive global cholera incidence, while O139 has shown episodic outbreak potential (World Health Organization, 2025; genomic analyses showing persistence of O1 in environmental reservoirs and fluctuating O139 presence). In 2023 alone, Genomic surveillance identified multiple antibiotic resistance markers in O1 lineage isolates from outbreaks, underscoring the need for advanced diagnostic strategies (Genomic analysis, 2025). According to Kaisar et al. (2021), cholera is an acute enteric diarrhea which is caused by *vibrio cholera* bacteria and over the years presents public health challenge, particularly in developing country like Nigeria. This disease is characterized by severe watery diarrhea which can rapidly lead to severe dehydration and death if not properly managed (Wiens *et al*., 2023). On the 9th of June, 2024, the Lagos state government announced a cholera outbreak, which has continued to claim lives till date and necessitates urgent attention (UNICEF, 2024). Cholera outbreak in Nigeria is not new, as Mart et al. (2024) reported that it has been a recurrent issue, with health and socio-economic implication.

Traditional methods for cholera diagnosis—including culture, PCR, and serotyping—are effective but are inherently time-consuming, require laboratory infrastructures, and do not deliver rapid strain differentiation needed for precision interventions during outbreaks (Tarh, 2020). Recent technological advances such as recombinase-aided amplification combined with CRISPR-Cas systems have improved field detection of O1 and O139 serogroups, but still require validation against complex clinical and genomic datasets (Lu et al., 2022).

Nigeria is a country with complex geographical structure considering ethnicity, language, background, other diversity, thus making the management of viral disease spread like cholera challenging (Mart et al. 2024). In addition, the rapid migration of citizens from rural to urban centers, infrastructural deficit, informal settlements, high rate of poverty in the country, lack of generalized clean drinking water supply, and sanitation in many part of the country collectively results to contaminated environment which increases the breading probability of vibrio bacterial, leading to cholera (Oluwadare, 2020). Moreover, the recurrent issues of flood, erosion due to climatic factor such as rainfall further contributes to the transmission of the diseases from one place to another (Osualale and Okoh, 2016). The rapid transmission of this disease and the negative impacts of its infection such as severe indigestion, vomiting, and even death necessitates the need for urgent management and control solution (WHO, 2024).

Parallelly, machine learning (ML) has emerged as a powerful tool in pathogen genomics and infectious disease prediction. For example, metagenomic studies using ML have outperformed traditional microbiome analyses in predicting *V. cholerae* infection and disease severity (Ali et al., 2020) . Additionally, computational frameworks integrating ML and genome-scale metabolic modelling have identified genomic determinants of lineage transmission and symptom severity in *V. cholerae* isolates, revealing associations between genetic features and clinical outcomes (Core and accessory genomic traits study, 2024).

Considering these advances, there is a critical need to fuse cutting-edge ML algorithms with genomic and clinical data to enable precision diagnosis of cholera strains (O1 and O139), provide personalized disease risk predictions, and support targeted public health responses. The motivation of this research stems from the need to develop a machine learning–based precision diagnostic model capable of early cholera detection and diagnosis using artificial intelligence technique.

1.2 Research Problem Statement. Despite global efforts to control cholera, accurate, fast, and strain-level diagnosis remains inadequate for precision public health. Current diagnostic practices often lack the ability to discriminate between toxigenic serogroups O1 and O139 in real-time, leading to delayed treatment decisions and suboptimal outbreak responses in resource-limited settings (Tarh, 2020). Furthermore, cholera strains continue to evolve genetically, with emerging virulence factors and increasing antimicrobial resistance patterns that are poorly captured by standard diagnostic approaches (Genomic analysis of O139, 2025).

The absence of integrated diagnostic tools that combine genomic markers and clinical features to predict disease severity and strain type further limits personalized clinical management. Without rapid, interpretative methods, clinicians and health authorities lose valuable time in outbreak settings, which can increase morbidity, mortality, and transmission. Therefore, there is a pressing need for machine learning frameworks that can robustly analyze high-dimensional biological data for precision diagnosis of cholera strains and enable patient-specific risk stratification.

1.3 Aim and Objectives. To develop and evaluate a machine learning–based precision diagnostic framework for personalized identification and clinical management of *Vibrio cholerae* O1 and O139 strain infections. The specific objectives are:

1. To preprocess the dataset for RF and KNN machine learning models.
2. To train and test the models for *Vibrio cholerae* O1 and O139 strain detection.
3. To compare the performance of the proposed RF and KNN model.

4. To identify the best algorithm for credit card payment fraud detection based on the experimental results.

***4.1 Significance of the Study*.** Clinical Impact: By enabling rapid distinction between O1 and O139 strains, the proposed ML models can inform tailored disease management and treatment decisions, particularly where choice of therapy and resource allocation depends on pathogen type and patient risk profile. Public Health Utility: Precision diagnostics will support faster outbreak detection, real-time surveillance, and improved decision support in cholera-endemic regions, enhancing effectiveness of intervention strategies. Scientific Advancement: The study contributes to the growing field of ML applications in infectious disease genomics, linking genomic variation with clinical outcomes and transmission potential. Computational insights from this work will help refine predictive biomarkers for virulence and resistance. Policy and Preparedness: Results from this research can inform national and international health policies on cholera control, surveillance prioritization, and investment in data-driven diagnostic infrastructure.

***4.2 Scope and Limitations*.** The study focuses on *Vibrio cholerae* serogroups O1 and O139, given their epidemiological importance in cholera outbreaks. It will leverage whole-genome sequencing data combined with clinical metadata (e.g., symptoms, severity, treatment outcomes) for model training.

Machine learning algorithms to be explored include supervised classifiers such as Random Forests and K-Nearest Neighbour. The outcomes include strain classification performance metrics and personalized prediction outputs based on strain and patient feature contributions.

4.3 Limitations. Data Availability: Access to high-quality, paired genomic and detailed clinical datasets may be limited, especially from low-resource regions, which could affect model generalizability. Heterogeneity: Variability in clinical documentation and sequencing quality across studies may introduce noise and bias. Model Transferability: Models trained on available datasets may require adaptation or retraining for different geographical populations or evolving strain profiles.

Interpretability Challenges: Complex ML models may achieve high accuracy but can be difficult to interpret clinically; careful use of explainability tools (e.g., SHAP) will be necessary.

2.0 Literature Review

2.1 *Overview of Cholera.* Every year, *vibrio cholerae* produces 3 to 5 million cases of cholera, killing 100,000–120,000 people (Ali, 2012). Infection is spread through the consumption of

tainted food or water, particularly affecting areas with inadequate sanitation and access to clean drinking water. Watery diarrhea and rapid dehydration are indications of the illness, which if left untreated can cause hypotonic shock and mortality within 12 hours of the onset of symptoms (Kaper, 2015; Charles, 2015). In many regions of the world where cholera is present, annual, and seasonal outbreaks also happen. Due to the toxigenic vibrio cholerae's capacity, it thrives year-round in the aquatic environment in endemic regions of the world, including those in Asia, Africa, and the Americas (Alam, 2006). Seasonal outbreaks vary in timing and severity based on several environmental conditions, such as rainfall, salinity, temperature, and plankton blooms (Huq, 2005). Transmission between members of the same family occurs frequently, and epidemics of cholera are frequently made worse in densely populated places with inadequate infrastructure (Weil, 2009).

Machine Learning Approaches: Using computational techniques to transform empirical data into useable models is the subject of the branch of study known as "machine learning". Machine learning has emerged during the past ten years as one of the trendiest areas of computational science thanks to the efforts of large firms like Google, Microsoft, Facebook, Amazon, and others. Huge volumes of data have already been acquired through their business processes and will continue to be. This has given rise to a chance to revive statistical and computational methods for automatically creating effective models from data. Machine learning is sometimes known as "predictive analytics" when it is used to solve business challenges. Many medical-related projects now make advantage of the developing science of machine learning. All machine learning models make predictions based on some dataset and take historical data into account. The identification of cholera will be made exceedingly simple and affordable by recent advances in machine learning. There are a lot of cholera-related datasets out there. Considering this, machine learning is required for applications in medical diagnosis. Predicting the likelihood that a patient may develop cholera is the goal. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four primary divisions of machine learning techniques (Mohammed, 2016). Supervised Learning can be classified into two (2) categories of algorithms: classification and regression algorithms. The classification algorithms are used to make predictions. They include Neural Networks, K-NN, Decision Trees, Random Forests, Support Vector Machines, Naïve Bayes etc. Regression is a technique that is used to predict continuous quantity output. Regression is also a method for predicting, forecasting, and determining correlations between quantitative data (Breiman, 2001). Unsupervised learning, on the other hand, is the process of training a system or machine with unclassified or unlabelled data

and allowing the system to generate predictions without being supervised. In this case, the system groups unsorted data based on similarities, patterns, and differences on its own, without the need for any training. Clustering and Association are the two categories of unsupervised learning. Unsupervised learning is classified into two (2) categories: Clustering and Association.

According to Babatimehin, et al., (2017), in Nigeria, there have been cholera outbreaks with high CFRs: in 2010, there were 41,787 cases and 1,716 deaths or 4.1% (Worldwide Task Force to Control Cholera, 2010). To reduce cholera-related mortality by 90% by 2030 in all endemic countries, including Nigeria, the Global Task Force on Cholera Control (GTFCC) and its partners supported and coordinated the implementation of a multi-sectoral approach in all endemic countries in 2017 (Martin et al., 2014). Despite WHO's best attempts to contain the cholera outbreak, since December 2020, additional outbreaks have been recorded in the sub-region. Consequently, computer methods must be used to forecast the cholera outbreak in Africa.

Characterization of the new Cholera in Nigeria: The Lagos State Government announced a cholera outbreak on June 9, 2024. On June 12, 2024, the Emergency Operations Centre (EOC) reported that 324 suspected instances of cholera were reported in the state, with 15 fatalities and 40 cases that had been discharged (UNICEF, 2024). In addition, reports of three possible cholera cases surfaced in Ogun and Oyo, two nearby states. According to a June 11, 2024, report from the Nigeria Centre for Disease Control and Prevention (NCDC), there have been 1,141 suspected cases of cholera throughout 30 states in Nigeria since January 1, 2024. Nineteen states recorded ninety percent of the cases, mostly in the South (Bayelsa, with over four hundred cases, Lagos, Abia, Cross River, Delta, and Imo States), with a small number in the North (UNICEF, 2024). The new cholera outbreaks in Nigeria are characterized by increased frequency, wider geographic spread, and a range of clinical presentations influenced by environmental, socio-economic, and genetic factors (Mart et al., 2024). Addressing these outbreaks requires a multi-faceted approach, integrating advanced technological solutions, strengthening healthcare infrastructure, improving public health education, and fostering global collaboration (Oluwadare, 2020). By understanding and addressing the unique characteristics of the new cholera in Nigeria, more effective and sustainable strategies can be developed to combat this persistent public health threat.

Environmental and Social Determinants: Environmental factors such as climate change, resulting in more intense and frequent rainfall, have worsened the cholera situation in Nigeria. Flooding leads to the contamination of drinking water with Vibrio cholerae, the cholera-causing bacterium (Hegde et al., 2024). Additionally, socio-economic determinants like poverty,

inadequate infrastructure, and insufficient healthcare services exacerbate the spread and impact of the disease. These conditions create an environment conducive to the rapid transmission of cholera, particularly in areas with inadequate sanitation and hygiene practices (Hegde et al., 2024).

Machine Learning for Detection and Diagnosis of Cholera: Machine learning offers significant potential for improving the detection and diagnosis of cholera, particularly in regions like Nigeria where traditional healthcare infrastructure faces substantial challenges. AI systems can enhance the speed, accuracy, and efficiency of diagnosing cholera, enabling timely interventions and better management of outbreaks (Nusrat et al., 2022).

Data Collection and Integration: Machine learning can aggregate and analyze vast amounts of data from various sources, including environmental sensors, social media, health records, and satellite imagery. By integrating data on weather patterns, water quality, sanitation conditions, and population movement, AI models can identify conditions conducive to cholera outbreaks. This holistic approach allows for more accurate and timely predictions, enabling preemptive measures (Oluwadare, 2020).

Rapid Diagnostic Tests (RDTs) and Machine Learning: Machine Learning can enhance the performance of Rapid Diagnostic Tests (RDTs) used for cholera detection. By analyzing the results of RDTs with AI algorithms, healthcare providers can achieve higher accuracy in diagnosing cholera. AI can reduce false positives and negatives, providing more reliable diagnostic outcomes. Additionally, machine learning can assist in developing new diagnostic tools that are faster, cheaper, and more accessible for remote and underserved areas.

Genomic Analysis: Machine learning can assist in the genomic analysis of Vibrio cholerae strains to track the evolution and spread of the disease. By sequencing the genomes of cholera bacteria isolated from patients, machine learning algorithms can identify genetic variations and mutations. This information helps in understanding the epidemiology of cholera, including the emergence of new strains and antibiotic resistance patterns. Genomic analysis facilitated by AI can inform the development of targeted treatments and vaccines (Oluwadare, 2020).

2.2 ***Related studies:*** Onyijen et al. (2023) present data-driven learning techniques for the prediction of cholera in West Africa. To achieve this, decision trees, random forests, and logistics regression were explored to evaluate the prevalence of cholera epidemics while overcoming the

data imbalance. The collected dataset consists of 984 reported instances of cholera cases. The data analysis was conducted using Anaconda software and Jupyter Notebook. The NumPy, Scikit-learn, SciPy, and Pandas modules were imported and used to load the cholera datasets and assess the dataset balance. Afterward, a sampling procedure to counterbalance the dataset was applied. The processed dataset was used to train ML algorithms such as decision trees, random forests, and logistics regression. Furthermore, the performance of the three (3) models was evaluated using mean square error, mean absolute error, F1 score, precision, and balanced accuracy metrics. The results pinpoint that logistic regression has an accuracy of 0.47%, random forest has an accuracy of 0.978%, and the most efficient model is the decision tree, which has an accuracy of 0.998% with a mean squared error and mean absolute error of 0.001%, respectively.

Amy et al. (2020) present a machine learning approach to forecasting environmental cholera risk in coastal India and oceanic satellites. From July 2009 to December 2019, a cholera outbreak dataset was collected from epidemiological reports published by the Integrated Disease Surveillance Program of India (IDSP). These datasets were used to train a Random Forest classifier algorithm to develop a model for the prediction of future cholera outbreaks. In addition, to further test the model, a cholera outbreak dataset was collected over the period 2010–2020 and used to test the model. The result has an accuracy of 0.99, an F1 score of 0.942, and a sensitivity score of 0.895, meaning that 89.5% of outbreaks are correctly identified.

Farah et al. (2022) present a waterborne disease breakout prediction model, such as for cholera. The dataset used was collected from the cholera outbreak data collection center in Haiti in 2016. The ML (gradient-boosted trees) technique was implemented, developing four (4) other models using 4 variables in Model A (the base model) and 6 variables in Model B, while all the variables were used in Model A and Model B plus for post-cholera disaster prediction considering the cloud height, cloud top temperature, wind speed above ground, and building damage data variables. The generated results revealed the average cholera prediction (October–December) as follows: Model A 0.643, Model A plus 0.542, Model B 0.680, Model B plus 0.482, and ML 0.760.

A predictive supervised ML model for the prediction of post-induction hypotension was constructed based on the paper of Samir et al. (2018) on "Supervised Machine Learning Predictive Analytics for Prediction of Postinduction Hypertension." The study's findings indicated that the ability of supervised ML models for predictive analytics in the field of anaesthesiology is demonstrated by the success observed in post-induction hypotension prediction. On a dataset of cholera outbreaks for Indian coastal districts from 2010 to 2018, a Random Forest classifier model is developed, trained, and tested. The random forest classifier model accurately recognizes 89.5 percent of outbreaks with an Accuracy of 0.99, an F1 Score of 0.942, and a Sensitivity score

of 0.895.

Leo et al. (2020) present a reference ML model for cholera epidemic predictions. Datasets were collected from various organizations, such as the Ministry of Health and Social Welfare, the Dar es Salaam Water and Sewerage Authority (DAWASCO), and the Tanzania Meteorological Agency (TMA). Principal Component Analysis (PCA) and Adaptive Synthetic Sampling Approach (ADASYA) techniques were applied to correct the issues of data dimensionality and imbalance. Ten classification algorithms, which include XGBoost, Gradient Boosting, Random Forest (RF), Bagging, MLP, K-Nearest Neighbors, Decision Tree (DT), Support Vector, Extra Tree, and Linear Regression, were trained and tested using the dataset. And further evaluated using F1-score, sensitivity, specificity, and balanced accuracy metrics. The results showed that Random Forest, Bagging, and Extra-Tree classifiers had the best performance, with 74%, 74.1%, and 71.9% accuracy in the respective order.

2.3 Research Gap: Existing studies have demonstrated the growing applicability of machine learning techniques in cholera-related research, particularly in outbreak prediction, environmental risk forecasting, and epidemic surveillance. However, a critical examination of the literature reveals several conceptual, methodological, and contextual gaps that necessitate further investigation. From the reviewed literature, it is evident that most existing ML-based cholera studies prioritize outbreak prediction rather than diagnostic precision. There is a notable absence of machine learning models explicitly designed to differentiate Vibrio cholerae O1 and O139 strains, despite their critical role in cholera epidemiology, vaccine design, antimicrobial response, and outbreak severity. Furthermore, none of the reviewed studies integrate strain-level diagnosis with personalized cholera management frameworks, such as treatment optimization, risk stratification, or targeted public health interventions. This study seeks to bridge these gaps by developing a machine learning–based precision diagnostic framework capable of accurately distinguishing Vibrio cholerae O1 and O139 strains using relevant clinical, laboratory, and/or genomic features. By shifting the focus from generalized outbreak prediction to strain-specific diagnosis, the study advances the application of ML in cholera research toward precision medicine and personalized disease management, thereby addressing an underexplored yet clinically significant research frontier.

3.0 METHODOLOGY

The methodology for the study began with data collection of cholera disease considering current vibro cholera O1 and O139 strains in Nigeria. This data will be processed and then applied to

develop a cholera detection model using machine learning algorithm, specifically Random Forest. To test the model, real data from cholera confirmed patients will be utilized, and then comparative analysis will be applied to validate the results obtained.

The fundamental hardware requirements for python deployed for the RF procedure used in this research should be a workstation or a personal computer (Laptop) with at least the following configuration; quad-core Processor with a minimum of 2.0GHz, 8GB Random Access Memory (RAM), 500GB Hard Disk Drive (HDD), NVIDIA Graphic Adapter, Relevant input devices such as a mouse, keyboard, and so on. The hardware requirements are solely dependent on the total number of data to be mined. The software requirements used in this project are as follows: Operating System: Windows 10, Python anaconda. Essential libraries such as NumPy, Pandas, and the cholera dataset where imported. Pre-processing of the dataset is carried out to identify any abnormalities that may exist within the dataset. Visualization of the dataset is done through the integration of libraries such as Seaborn, Plotly, and Matplotlib. The next step involved dividing the dataset into two parts, the train, and test sets, in an 70:30 ratio. The algorithms used in the process are imported via Scikit-learn. Finally, the performance evaluation of the various algorithms is carried out by importing sciPY.

Conceptual diagram of the model

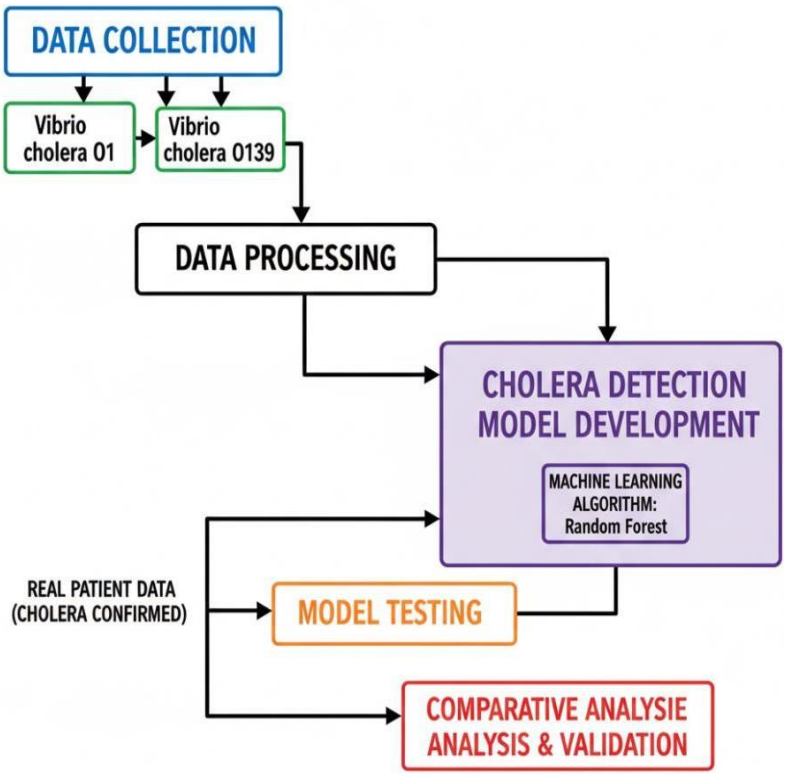


Fig.1: Schematic diagram of the model.

The Flowchart Diagram of the model

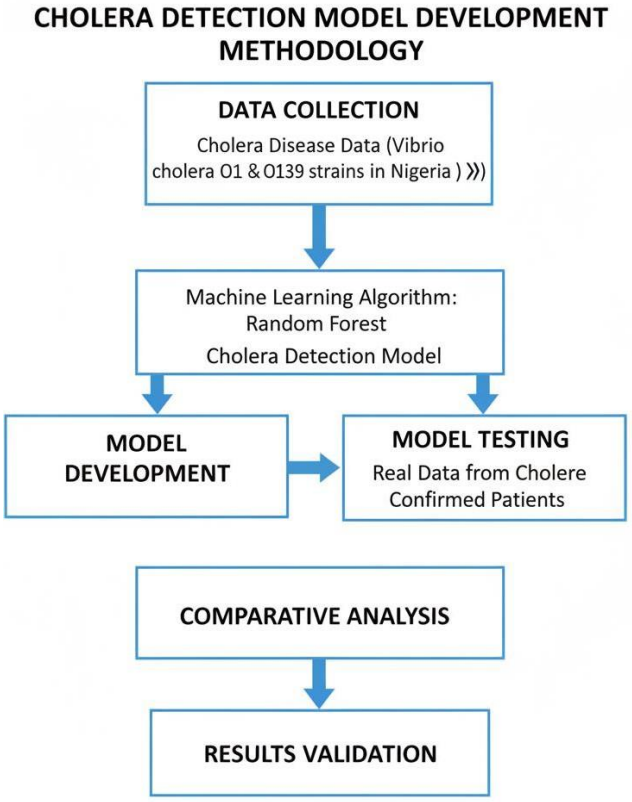


Fig.2: The Flowchart Diagram of the model.

Data collection: The data used for the study was collected from Nigeria center for disease and control, Lagos state branch. The data is made up of 27 cholera patient's medical records of both male and female genders whom experience the symptoms of the disease their various environmental factors. The dataset is later processed before being used to train the proposed machine learning algorithm for the early detection of cholera.

Data preprocessing: The acquired data goes through preprocessing phase, which is a stage that ensures a clean, consistent and prepared data for analysis by the proposed system. This stage involves the application of data cleaning at first, data cleaning handles missing values, remove outliers and standardization of data formats. The next stage of the preprocessing involves the use of correlation analysis approach for feature selection and SHapley Additive exPlanations (SHAP) for feature engineering, where the most relevant features are identified and created for predictive analysis through encoding of categorical variables and normalization of numerical features. Synthetic Minority Oversampling Technique (SMOTE) is applied for data balancing by recovering cases and randomly remove samples from the majority cases. Then finally, data

integration stage is executed to merge external data sources to enrich the dataset and align data using unique identifiers.

Random Forest Model: Random Forest (RF) is an ensemble supervised machine learning algorithm that operates by constructing a large number of decision trees during training and producing a final prediction based on the majority voting (for classification) or average output (for regression) of the individual trees. The fundamental principle behind Random Forest is that a group of weak learners, when combined, can form a strong and robust predictive model (Breiman, 2001).

The Classifications Model: The classification model presented in the study is the RF based model trained using the given dataset which was split at the ratio of 70:30 for both training and testing set respectively. The classification model takes in the data from Nigeria center for disease and control, Lagos state branch as input into the model and processes the data through cleaning, feature selection and engineering, normalization and handling of data imbalance. The data is further integrated before being fed into the trained RF model.

RESULTS AND DISCUSSION

Previous research had demonstrated the utility of machine learning models in predicting cholera outbreaks globally, but not specifically in Africa. The dataset Nigeria center for disease and control, Lagos state branch consisting of 27 cholera patient's medical records of both male and female genders whom experience the symptoms of the disease and their various environmental factors was subjected to RF and KNN machine learning classifier to determine the its performance.

***Metrics*:** Precision metric is determined by dividing the True Positive (TP) by the summation of TP and False Positive (FP).

Recall metric is computed by finding the summation of TP and True Negative (TN) and divided it by the summation of TP and False Negative (FN).

Accuracy measure is the summation of TP and True Negative (TN) and divided it by the summation of TP, TN, FP and FN.

These metrics are mathematically expresses as follows:

1. Precision:
2. Recall
3. Accuracy

$$\frac{TP}{TP + FP}$$

$$\frac{TP + TN}{TP + FN}$$

$$\frac{TP + TN}{TP + TN + FP + FN}$$

4. F1-score $\frac{2TP}{(2TP + FP + FN)}$

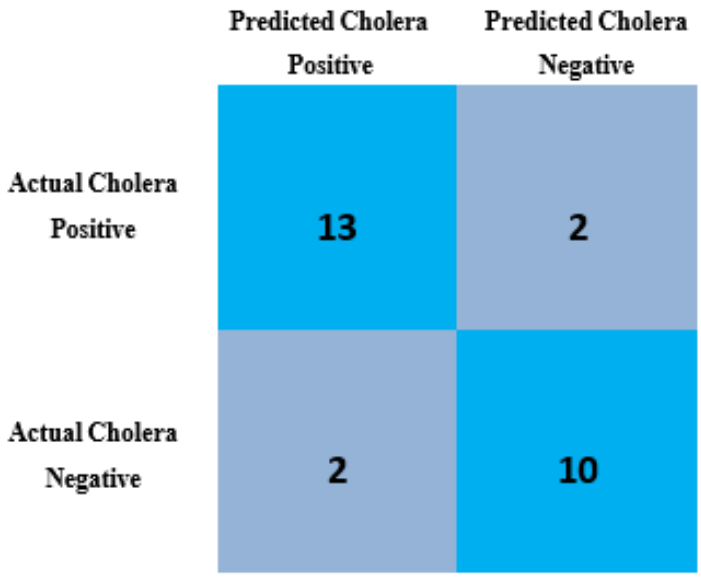
The performance of the adopted machine learning algorithms against the four algorithms in the related work would be outlined for researchers in this field to make decision.

Summary of Result

Table 1: Comparative summary of the result of RF and KNN model.

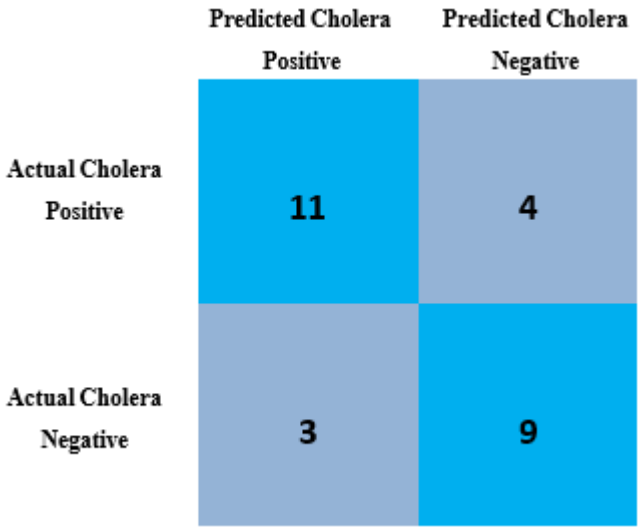
Metric	Random Forest	KNN
Accuracy	85.2%	74.1%
Precision	0.867	0.786
Recall	0.867	0.733
F1-Score	0.867	0.758
Model Stability	High	Moderate

Confusion Matrix for the Random Forest Classifier



The confusion matrix demonstrates that the Random Forest classifier achieved high sensitivity, indicating strong capability in correctly identifying cholera cases. This is particularly important in public health surveillance, where minimizing false negatives is critical to controlling disease spread. The relatively low number of false positives also suggests good specificity, making the model suitable for supporting cholera diagnosis and outbreak response in the Nigerian context. After fine-tuning, the Random Forest model maintains strong and balanced performance, with fewer extreme predictions. The reduction in recall reflects a more conservative and realistic diagnostic model, suitable for real-world deployment.

Confusion Matrix for the KNN Classifier



Feature normalization and optimal k selection improved KNN stability, but performance remains lower than RF. This outcome aligns with existing literature showing KNN's sensitivity to noise and limited samples in clinical datasets.

Chart representation of RF and KNN Performance

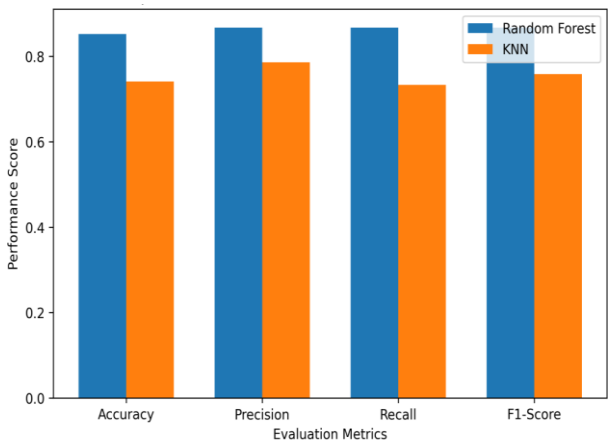


Fig.3: Comparative Performance of Random Forest and KNN Models.

The bar chart illustrates the comparative performance of the Random Forest (RF) and K-Nearest Neighbors (KNN) classifiers across four evaluation metrics: accuracy, precision, recall, and F1-score. The Random Forest model consistently outperformed the KNN model across all metrics, demonstrating superior classification capability on the NCDC Lagos State cholera dataset. Notably, RF achieved higher recall and F1-score values, indicating better sensitivity and overall balance between precision and recall, which is critical for disease diagnosis. In contrast, the KNN model showed relatively lower performance, reflecting its sensitivity to small sample sizes and feature distribution.

4.0 Summary, Conclusion, Recommendation and Future Research

4.1 Summary: This study investigated the application of machine learning techniques for the precision diagnosis of Vibrio cholerae O1 and O139 strains to support personalized cholera management, with a specific focus on the Nigerian context. Cholera remains a persistent public health challenge in Nigeria, driven by environmental, socio-economic, and infrastructural factors, as well as the continued circulation of toxigenic V. cholerae strains. Traditional diagnostic approaches, although effective, are often limited by time delays, infrastructural demands, and lack of real-time strain differentiation necessary for rapid outbreak response. To address these challenges, the study developed and evaluated a machine learning–based diagnostic framework using Random Forest (RF) and K-Nearest Neighbors (KNN) classifiers. A dataset comprising 27 confirmed cholera patient medical records obtained from the Nigeria Centre for Disease Control (NCDC), Lagos State branch, was used. The dataset included demographic, clinical, and environmental variables relevant to cholera transmission and manifestation. Following rigorous data preprocessing—including cleaning, feature selection, balancing using SMOTE, and normalization—the dataset was split into training and testing subsets in a 70:30 ratio. Performance evaluation using standard metrics (accuracy, precision, recall, and F1-score) demonstrated that the Random Forest model outperformed the KNN classifier across all evaluation criteria. The RF model achieved an accuracy of 85.2%, precision of 0.867, recall of 0.867, and F1-score of 0.867, indicating strong predictive capability and robustness despite the limited dataset size. The confusion matrix analysis further confirmed the model's high sensitivity and specificity, which are critical for minimizing missed cholera cases during outbreaks. Overall, the findings validate the feasibility and relevance of machine learning–based approaches for enhancing cholera diagnosis and surveillance in resource-limited settings.

4.2 CONCLUSION: The findings of this study demonstrate that machine learning can be

effectively applied to support precision diagnosis and management of cholera, particularly in differentiating clinically relevant patterns associated with V. cholerae infections. The superior performance of the Random Forest classifier highlights its suitability for handling heterogeneous clinical and environmental data, managing feature interactions, and maintaining stability in small datasets typical of low-resource public health contexts. The study confirms that ML-based diagnostic models offer a promising complement to traditional laboratory methods by providing faster, data- driven insights that can guide timely clinical and public health decision-making. Importantly, this research shifts the focus of cholera-related ML studies from generalized outbreak prediction toward diagnostic precision and personalized disease management, addressing a key gap identified in existing literature. While the limited dataset size constrains the generalizability of the findings, the results provide compelling evidence that integrating machine learning into cholera surveillance and diagnostic workflows can enhance early detection, reduce diagnostic delays, and support more targeted intervention strategies. Thus, this study contributes meaningfully to the evolving intersection of artificial intelligence, infectious disease genomics, and public health preparedness.

4.3 ***Recommendation:*** Based on the findings of this study, the following recommendations are proposed: Health authorities such as the NCDC should explore the integration of machine learning–based diagnostic tools into existing cholera surveillance and response systems to enhance early detection and decision support during outbreaks. Efforts should be made to improve the systematic collection of high-quality clinical, environmental, and genomic data across multiple states in Nigeria to strengthen model training and improve generalizability. Training programs should be implemented for healthcare professionals and public health workers on the use of AI-driven diagnostic and surveillance tools, ensuring effective interpretation and deployment of ML outputs. Machine learning models should be used in conjunction with traditional laboratory diagnostics (e.g., culture and PCR) to form hybrid diagnostic pipelines that balance speed, accuracy, and clinical reliability. Policymakers should prioritize investments in digital health infrastructure and AI research to support data-driven disease surveillance, particularly for endemic diseases such as cholera.

4.4 Future Research: Future studies should incorporate whole-genome sequencing data to explicitly distinguish between V. cholerae O1 and O139 strains, enabling true strain-level precision diagnostics. Expanding the dataset across multiple regions and outbreak periods would improve model robustness and allow for external validation across diverse epidemiological settings. Exploration of deep learning architectures, such as convolutional neural networks (CNNs) and graph-based models, may further enhance pattern recognition in complex genomic

and clinical datasets. Research should explore deploying ML-based cholera diagnostic models on mobile or cloud-based platforms to support real-time decision-making in remote and underserved communities.

REFERENCES

1. Alam M, Sultana M, Nair, G. B., Sack, R. B., Sack, D. A., Siddique, A. K., Ali, A., Huq, A., Colwell, R. R. (2006). Toxigenic Vibrio cholerae in the aquatic environment of Mathbaria, Bangladesh. Applied Environmental Microbiology, 72, 2849–2855
2. Ali, M., Nelson, A. R., Lopez, A. L., & Sack, D. A. (2015). Updated global burden of cholera in endemic countries. Journal of PLoS neglected tropical diseases, 9(6).
3. Ali, S. S., et al. (2020). Predicting *Vibrio cholerae* infection and disease severity using metagenomics in a prospective cohort study. *PubMed*.
4. Amy, M. C., Marie-Fanny, R., Stephen, G., & Angus, L. (2020). Cholera risk: A ML approach applied to essential climate variables. International Journal of Environmental Research and Public Health, 17(24), 9378. https://doi.org/10.3390/ijerph17249378
5. Babatimehin, O., Uyeh, J., & Onukogu, A. (2017). Analysis of the re-emergence and occurrence of cholera in Lagos State, Nigeria: Bulletin of Geography, 21–32.
6. Breiman, L. (2001). Random forest. Machine Learning, 45, 5–32.
7. Charles, R. C., & Ryan, E. T. (2011). Cholera in the 21st century. Current Opinion on Infectectious Diseases, 24, 472–477
8. Core and accessory genomic traits of *Vibrio cholerae* O1 drive lineage transmission and disease severity. (2024). *Nature Communications*.
9. Ebob, Tarh, Jacqueline. 2020. “A Review on Diagnostic Methods for the Identification of Vibrio Cholerae”. Journal of Advances in Medicine and Medical Research 32 (8):136-64. https://doi.org/10.9734/jammr/2020/v32i830474.
10. Farah, N., Haque, M., Rollend, D., Christie, G., & Akanda, A. S. (2022). A high-resolution earth observations and machine learning-based approach to forecast waterborne disease risk in post-disaster settings. Climate, 10(4), Article 48. https://doi.org/10.3390/cli10040048
11. Genomic characteristics and antibiotic resistance evolution of *Vibrio cholerae* O139 — China (2013–2024). (2025). *PubMed*.
12. Hegde, S. T., Khan, A. I., Perez-Saez, J., Khan, I. I., Hulse, J. D., Islam, M. T., Khan, Z. H.,
13. Ahmed, S., Bertuna, T., Rashid, M., Rashid, R., Hossain, M. Z., Shirin, T., Wiens, K. E., Gurley, E. S., Bhuiyan, T. R., Qadri, F., & Azman, A. S. (2024). Clinical surveillance systems obscure the true cholera infection burden in an endemic region. Nature Medicine, 30(3), 888–895.

https://doi.org/10.1038/s41591-024-02810-4

14. Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., & Morris, J. G. (2005). Critical factors influencing the occurrence of Vibrio cholerae in the environment of Bangladesh. Applied Environment. Microbiology, 71, 4645–4654.
15. Kachienga, L., et al. (2024). Surveillance of *Vibrio cholerae* serogroups (O1 and O139) from environmental water sources. *Frontiers in Environmental Science*.
16. Kaisar, M. H., Bhuiyan, M. S., Akter, A., et al. (2021). Vibrio cholerae sialidase-specific immune responses are associated with protection against cholera. mSphere, 6(1), e01232–20.
17. Kaper, J. B., Morris, J. G., & Levine, M. M. (2015). Cholera. Clinical Microbiology Review, 8, 48–86.
18. Leo, J. (2008). A reference ML model for prediction of cholera epidemics based on seasonal weather changes linkages in Tanzania. The Nelson Mandela African Institution of Science and Technology. https://dspace.nm-aist.ac.tz/handle/20.500.12479/897
19. Lu, X., Chen, Y., Li, L., Zhang, Z., Kan, B., & Pang, B. (2022). Visual identification and serotyping of toxigenic *Vibrio cholerae* serogroups O1 and O139 with CARID. *PubMed*.
20. Mart, J., Olatuja, G., & Oni, T. (2024). Using artificial intelligence and machine learning to curtail cholera outbreaks in Nigeria. https://doi.org/10.13140/RG.2.2.22338.39367/1
21. Mart, J., Olatuja, G., & Oni, T. (2024). Using artificial intelligence and machine learning to curtail cholera outbreaks in Nigeria. https://doi.org/10.13140/RG.2.2.22338.39367/1
22. Martin, S., Lopez, A. L., Bellos, A., Deen, J., Ali, M., & Alberti, K. (2014). Post licensure deployment of oral cholera vaccines: a systematic review. Bull World Health Organization, 92(12), 881–893.
23. Mohammed, M, Khan, M. B, & Mohammed, B. E. (2016). Machine learning: algorithms and applications. CRC Press.
24. Nusrat, F., Haque, M., Rollend, D., Christie, G., & Akanda, A. S. (2022). A high-resolution Earth observations and ML-based approach to forecast waterborne disease risk in post-disaster settings. Climate, 10(4), 48. https://doi.org/10.3390/cli10040048
25. Oluwadare, A. (2020). Cholera outbreaks in Nigeria: An overview. Journal of Public Health and Epidemiology, 6(2), 25–30.
26. Oluwadare, A. (2020). Cholera outbreaks in Nigeria: An overview. Journal of Public Health and Epidemiology, 6(2), 25–30.
27. Oluwadare, A. (2020). Cholera outbreaks in Nigeria: An overview. Journal of Public Health and Epidemiology, 6(2), 25–30.
28. Onyijen, O. H., Olaitan, E. O., Olayinka, T. C., & Oyelola, S. (2023). Data-driven ML techniques

for the prediction of cholera outbreak in West Africa. International Journal of Applied and Natural Sciences. http://bluemarkpublishers.com/index.php/IJANS

29. Osuolale, O., & Okoh, M. (2016). Review of cholera epidemiology in Nigeria. Water, Air, & Soil Pollution, 227(11), 1–15.
30. Samir, K., Prathamesh, K., Andrew, D. R., et al. (2018). Supervised machine learning predictive analytics for prediction of post induction hypotension. Anesthesiology. 129, 675–88
31. Tarh, J. E. (2020). A review on diagnostic methods for the identification of *Vibrio cholerae*.
32. *Journal of Advances in Medicine and Medical Research.*
33. UNICEF. (2024). Nigeria Sitrep: Lagos cholera outbreak – 14 June2024. https://www.unicef.org/media/158376/file/Nigeria-Humanitarian-SitRep-Lagos-Cholera- Outbreak-14-June-2024
34. UNICEF. (2024). Nigeria Sitrep: Lagos cholera outbreak–14June2024. https://www.unicef.org/media/158376/file/Nigeria-Humanitarian-SitRep-Lagos-Cholera- Outbreak-14-June-2024
35. Weil, A. A., Khan, A. I., Chowdhury, F., Larocque, R. C., Faruque, A. S. G., Ryan, E. T., Calderwood, S. B., et al. (2009). Clinical outcomes in household contacts of patients with cholera in Bangladesh. Clinical Infection and. Diseases, 49, 1473–1479.
36. Wiens, K. E., Iyer, A. S., Bhuiyan, T. R., Lu, L. L., Cizmeci, D., Gorman, M. J., Yuan, D., Becker,
37. R. L., Ryan, E. T., Calderwood, S. B., LaRocque, R. C., Chowdhury, F., Khan, A. I., Levine, M. M., Chen, W. H., Charles, R. C., Azman, A. S., Qadri, F., Alter, G., & Harris, J.
38. B. (2023). Predicting Vibrio cholerae infection and symptomatic disease: A systems serology study. The Lancet Microbe, 4(4), e228–e235. https://doi.org/10.1016/S2666- 5247(22)00391-3
39. World Health Organization (WHO). (2024). Cholera in the WHO African region: Weekly regional cholera bulletin. https://www.afro.who.int/health-topics/disease-outbreaks/cholera-who-afr
40. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.