International Journal of Scientific Research in Modern Science and Technology
ISSN: 2583 -7605 (Online)
© IJSRMST | Vol. 3 | Issue 10 | October 2024
Available online at: https://ijsrmst.com/
DOI: https://doi.org/10.59828/ijsrmst.v3i10.255

Recent Advances in Machine-Learning Driven Cholera Research: A Review

Jessica Nwobodo[1], Shugaba Wuta[2], Michael Ibitoye[3], Paul Omagbemi,[4] Martins Offie[5]

[1] Technology Department, iNterra Networks, Federal Capital Territory, Abuja, Nigeria
jnwobodo@interranetworks.com
[2] School of Information and Technology, American University of Nigeria, Yola, Adamawa, Nigeria
shugaba.wuta@aun.edu.ng
[3] Electrical Electronics Engineering Department, Federal University of Technology, Minna, Minna, Niger State, Nigeria
Ibitoye.m1601575@st.futminna.edu.ng
[4]Computer Engineering Department, Federal University of Technology, Minna, Minna, Niger State, Nigeria
omagbemi.m1600484@st.futminna.edu.ng
[5]Computer Science Department, African University of Science and Technology, Federal Capital Territory, Abuja, Nigeria
okasiemobi@aust.edu.ng

Abstract

Cholera is a potentially epidemic and life-threatening secretory diarrheal disease caused by Vibrio cholerae, it is transmitted through the consumption of contaminated water. Cholera is prevalent in developing countries, characterized by inadequate access to clean water, sanitation and proper hygiene. Various studies have been conducted to evaluate its impact, predict its outbreak, and determine the best response during an epidemic. In conducting those studies, traditional mathematical and statistical models have been utilized, but more recently, artificial intelligence machine learning (ML) models have been used to better understand cholera, and eradicate it. AI models have shown higher accuracy than existing approaches, but there is no consolidated review study, comparing all the approaches used so far in literature. This is a review of the applications of ML techniques used in predicting, monitoring and preventing cholera epidemic. We analysed all relevant and recent studies that made use of various ML algorithms to address challenges in cholera prediction, diagnosis, transmission, and prevention. The reviewed works included the creation of models for early warning systems using environmental and socioeconomic data; improved diagnosis for more accuracy through the analysis of clinical data and serological markers; enhanced understanding of Vibrio cholerae genomics and evolution; and new approaches to water quality monitoring and cholera detection. ML methods have shown tremendous success in multiple areas, but not without challenges that still plagues further applications. Particularly, data unavailability, high data imbalance and real-time data collection. This review also highlighted areas of further application of ML such as real-time prediction of cholera outbreaks and region-agnostic models.

Keywords: Cholera, Machine Learning, Disease transmission, Public health crisis.

1. Introduction

Cholera is an acute diarrheal illness caused by infection of the intestine with Vibrio cholerae (*V. Cholerae*). It has an indirect transmission, where people can get sick from ingesting contaminated food or water (Fung, 2014). Diarrheal diseases were the fourth leading cause of death in Cameroon, with 50.4 and 41.4 deaths per 100,000 population from men and women, respectively. Since 1817, the world has faced seven cholera pandemics The first six pandemics of cholera induced millions of deaths across all continents. In spite of having virtually eradicated from affluent nations more than a century ago, cholera is still a major cause of disease and mortality in Africa, where it still affects a number of nations. The largest cholera outbreak to hit Malawi in 20 years is currently occurring, and Mozambique and Zambia, two nearby nations, have also reported a sizable number of cases. In the midst of a severe and protracted drought that has left millions of people in urgent need of humanitarian aid, other nations like Ethiopia, Kenya, Somalia, Burundi, Cameroon, the Democratic Republic of the Congo, and Nigeria are all currently dealing with cholera outbreaks. People who are already under stress from poverty, non-resilient health systems, violence, and poor infrastructure are disproportionately affected by cholera (Lawal *et al.*, 2024).

Fig 1. Shows the spread of cholera across African countries as it spans across west, eastern, and southern part of Africa. The dark green shaded countries highlight affected region.

Fig 1. Cholera Distribution Across Africa.

Machine Learning methods have been used to study cholera epidemics, in favor of mathematical models (Alfred & Obit, 2021). Supervised ML algorithms, particularly, classifiers are well more suited for forecasting cholera outbreaks (Ashari *et al.*, 2013). The use of ML to predict cholera has not been without challenges: The unavailability of balanced dataset, excessive data featuring, and the unavailability of real-time data collection are challenges faced with using ML for cholera prediction. Although these limitations exist, methods such as Synthetic Minority Oversampling Technique (SMOTE) based on nearest neighbor information, Adaptive Synthetic Sampling (ADASYN) have utilized to balance the data. Principal

Component Analysis (PCA) have been used to reduce the dimensions of the data, thereby, solving the challenge of excessive features.

In the following sections, this paper will analyze the contributions of ML in predicting, diagnosis, evaluating prevention measures against cholera using various classification algorithms.

2. Cholera: A Global Persistence and Regional Impact

Nuhu (2021) used the Naive Bayesian Classification with reported cholera data set in Yobe, Nigeria. The naïve bayes classifier was trained and tested with a total of 1443 cholera data from Yobe, Nigeria. The naïve bayes has the advantage of not skewing result based on missing data when the features used are independent. The resulting model had a training-set accuracy score of 99.49%, and a test-set accuracy of 99.41%. Confusion matrix resulted in 1 *Type I* error and 1 *Type II* error from a test of 337 data. The model showed that using more features like leg cramps, house condition, poor sanitation, lethargy, contaminated water, fatigue, fishy odor stool, dry mucous membranes, watery diarrhea, irritability, thirst, muscle cramps, rice water stool, loss of skin elasticity, rapid heart rate, vomiting resulted in a more accurate model. However, this research was not clear about the data collection of the extra features, nor explicit about what the features of the data indirectly collected from the Yobe state Ministry of Health. Also, the research acknowledged data imbalance and missing values in the dataset but did not mention the methodologies used to balance the data.

Campbell *et al.* (2020) proposed a Machine Learning approach applied to essential climate variables. The study takes a novel approach in utilizing machine learning to analyze Satellite-derived ECVs (Essential Climate Variables) to predict environmental cholera risk in India's coastal districts in 2010-2018, using a Random Forest Classifier to train and test the dataset. They found RFC effective for predicting cholera outbreaks using the ECVs and their lagged values. Analysis of the specific contribution of each pertinent climate variable to the model output revealed that *chlorophyll-a* concentration, sea surface salinity and land surface temperature are the strongest predictors of cholera outbreaks in the dataset used. However, the study was limited as the model performed better in areas with routine or annual cholera outbreaks but was less effective in areas with fewer outbreaks. The model was less effective in detecting sporadic, epidemic outbreaks of cholera which are more likely to occur due to external factors like contaminated travelers or natural disasters. The study was centered on the coastal district of India. This would constrain the model's generalizability, particularly in regions with distinct climatological and environmental variables.

Asadgol *et al.* (2019) proposed a study on the effect of climate change on cholera disease using an artificial neural network to evaluate the association between cholera cases and climate variables, to determine the most effective parameter. The authors collected daily records of cholera cases in Qom city from 1998 to 2016, along with observed climate variables such as maximum and minimum temperatures, and precipitation. Using Gamma Test (GT), the authors determined the best lag time and combination of input variables for the model, streamlining the selection process. Feedforward Multilayer Perceptron (MLP) type Artificial Neural Network (ANN) was adopted to simulate the impact of climate change on cholera incidence, using observed climate variables as predictors and cholera cases as the target outcome. The study

found cholera outbreaks and diarrhoea diseases are linked to climatic variables and climatic changes such as high temperature and low rainfall during dry seasons as bacteria thrive in such environments, thus, increasing the risk of cholera outbreak. Using statistical downscaling with LARS-WG, future climate projections for 2050 under scenarios RCP2.6 and RCP8.5 showed a trend toward warmer and wetter conditions. The authors highlighted their limitation as long-term detailed monitoring data on cholera and diarrhoea on patients' demographics such as age, sex, and socio-economic status. This limits the depth of the study's analyses. The findings are specific to Qom, Iran, indicating that the model may not directly apply to other regions with different climatic, social, and environmental conditions without further localised studies. Fig 2. Shows the Confusion matrix of the random forest model cholera outbreak prediction results on unseen test data

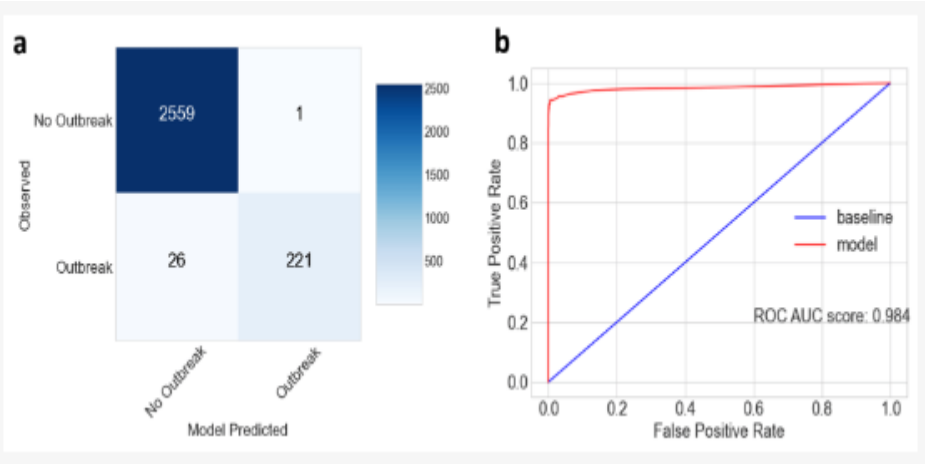

Fig 2.0 Cholera outbreak confusion matrix using the random forest model

Data imbalance and excessive features are purported to be responsible for low accuracy in cholera prediction models, adding complexity to including climate features that correlate with cholera epidemics. Nonetheless, Leo *et al.* (2019) trained models that combined reported cholera incidences with weather data in Tanzania. They used Adaptive Synthetic Sampling Approach (ADASYN) to balance the data, and Principal Component Analysis (PCA) to reduce dimensionality. The dataset was further used to train 7 ML algorithms: Extreme Gradient Boost (XGBoost), K-Nearest Neighbor (K-NN), Decision Trees (DT), Random Forest (RF), Extra Trees, AdaBoost and Linear Discriminant Analysis (LDA). Copies of the data with all combination of the data transformations were used to evaluate the 7 models. The base model for comparison were the models trained with the untransformed data. The results were evaluated using the sensitivity and specificity scores of the models. Across models, the specificity score dropped while using the oversampled data, while at the same it demonstrated improvement to the sensitivity data and the data imbalance. Overall, the XGBoost and K-NN have shown high accuracy, but XGBoost was chosen as the primary model because of its robustness, ability to use imbalanced data and speed. These crucial factors made the XGBoost better suited than K-NN.

Similar to this, Amshi *et al.* (2024) assessed the data transformation's performance improvement using the following methods: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to eliminate outliers, Synthetic Minority Oversampling Technique (SMOTE) for data class balancing, and Nonnegative Matrix Factorization (NMF) to resolve dimensionality reduction. Random Forest (RF), Naïve Bayesian (NB), and XGBoost were the categorization models that were employed.The models trained with the altered dataset shown notable gains over the models trained with the untransformed dataset, according to evaluations of the models using Matthew's correlation, Area Under Curve (AUC), accuracy, specificity, and sensitivity. Among all trained models and current cholera prediction models, the XGBoost also had the highest ranking (Campbell *et al.*, 2020; Leo *et al.*, 2019). Future research will focus on integrating real-time data to more accurately predict cholera.

Nusrat *et al.* (2022) proposed a model to identify and determine the impact of the environmental and social variables that play a significant role in post-disaster cholera outbreaks. Using the spatio-temporal satellite Earth Observation (EO) and the corresponding number of reported cholera reports in Haiti, after Hurricane Matthew in October 2016. The authors developed 4 geospatial models and a ML model using additive tree learning approach with a gradient boost algorithm; to predict the number of cases in a commune of Haiti (a commune consists of 140 communes for each of the 10 provinces). According to the following specification, the geographic models A, A Plus, B, and B Plus were produced: Four fundamental socio-environmental factors were used in Model A, which was developed using higher-resolution data than the earlier study. During the hurricane, Model B introduced extreme event variables. Cloud height, cloud top temperature, wind speed, and building damage statistics are all included in each Plus model. The information was obtained between July and December of 2016. The ML model outperformed the geographical model (Model B) in terms of performance.

Table 1.

Reference	**Data Transformation(s)**	**Accuracy Scores**	**Dataset**	**ML Algorithm**	**Data**
Nuhu (2021)	N/A	MCC=0.97689 Accuracy=0.9940 Specificity=0.9804 Sensitivity=0.9965	Cholera dataset from Yobe Ministry of Health	Naïve Bayesian (NB)	Leg spasms, unhealthy living conditions, tainted water, exhaustion, fishy-smelling feces,dry mucous membranes, watery diarrhea, Thirst, cramping in the muscles, rice-water stools, irritability, loss of skin suppleness, elevated heart rate, and vomiting
Campbell *et al.* (2020)	SMOTE	Accuracy=0.99 F1 Score = 0.942 sensitivity = 0.895	Surveillance data of cholera outbreaks (n=630; 2009-2019); Six Essential Climate	Random Forest (RF)	

			Variables (ECV); Satellite altimetry data product for the period 2016–2018		
Asadgol *et al.* (2019)	Gamma Test		Daily cholera prevalence data from January 1998 to December 2016 were gathered, while meteorological data for the period of 1976 to 2016 were obtained from the Iran Meteorological Organization (IMO).		
Amshi *et al.* (2024)	PCA + ADASYN	Sensitivity=0.805 ±0.169 Specificity=0.73±0 .05 Balanced accuracy=0.767±0. 09	Cholera cases and meteorological data were reported in the Dar es Salaam region from January 2015 to December 2017.	*XGBoost, K-NN, LDA, DT, RF, ExtraTrees, AdaBoost.	The data includes temperature, rainfall, humidity, wind, district location of the patient, the date of onset for cholera diagnosis, and patients' laboratory results.
Amshi *et al.* (2024)	NMF + SMOTE + DBSCAN	Matthew's correlation, area under curve (AUC), accuracy, specificity, and sensitivity	Cholera dataset from Yobe Ministry of Health Meteorological data from World Bank Climate Knowledge Socio-economic data from UNICEF Data Warehouse	*XGBoost, Naïve Bayesian (NB), Randomm Forest (RF)	
Nusrat *et al.* (2022)	-	-	-	Gradient Boost	Cloud temperature, water vapor, cloud weight, land surface temperature, cloud height, temperature anomaly, vegetation, land surface temperature anomaly, aerosol optical thickness, mean home address damage, cloud fraction, chlorophyl concentration, net radiation, sea surface temperature
Charnley *et al.* (2022)	Random forest variable importance		Cholera data were obtained from NCDC from 2018 - 2019	RF (regression model)	conflict, drought, Internally Displaced Persons, WASH, healthcare, population and poverty.
Onyijen *et al.* (2023)	-	Accuracy=0.998	Cholera case from west African countries 1970-2016	DT	

Charnley *et al.* (2022) introduced a machine learning approach to explore the impact of social and environmental extremes on cholera transmission in Nigeria. To understand cholera epidemic dynamics, the study examined variables such as population density, poverty, access to healthcare, water, sanitation, and hygiene (WASH) facilities, violence, drought, and internally displaced persons (IDPs). The Random Forest (RF) model was used to assess the covariates with the greatest impact through variable importance analysis. The reproduction number (R), a continuous parameter, was predicted by training an RF regression model using the k-10 fold method. A traffic light system was used to describe cholera outbreak triggers, with red, amber, and green indicating different R values ($R > 1$, $R < 1$). The study ascertained that alleviating poverty and enhancing access to sanitation facilities mitigated susceptibility to heightened cholera risk induced by extreme events such as monthly conflicts and fluctuations in the Palmer Drought Severity Index.

Onyijen *et al.* (2023) utilized the Random Forest to train and test the dataset. The random forest showed an accuracy value of 98% with a mean absolute error of 0.124 and a mean square error 5.952. The data presented showed the distribution of cholera cases in west Africa. This study can be improved by using an alternative algorithm to ascertain the accuracy rate. It is imperative to note that the accuracy can be further augmented by obtaining additional datasets from a plethora of diverse case studies.

3. Cholera Incidence, Genomics, and Water Safety

Machine learning models have been developed to gain a deeper understanding of how cholera interacts with the human body. For example, Azman *et al.* (2019) worked on improving the evaluation of cholera incidence in individuals. Cholera incidence is traditionally estimated based on reported cholera cases (Dick *et al.*, 2012). However, it is known that cholera cases are underreported (Charnley *et al.*, 2022). Azman *et al.* (2019) utilized serology data collected from 2006 to 2015 from cholera patients and their contacts to train a model that can detect whether an individual has been infected with cholera. The model employed confirmed cholera-negative serology markers from the close contacts of cholera patients as background data. The serology data of both infected and uninfected individuals were labeled and tracked over a period (up to 900 days) to provide more insight into the progression of infection for the model. The Random Forest (RF) model used single-marker thresholds and performed a 20-fold cross-validated area under the curve (cvAUC). The resulting model estimated the number of cholera incidences an individual had from day 0 to either day 10, 45, 100, 200, or 365. The model demonstrated high accuracy during testing and was externally validated. The external validation was performed using six months of sero-surveillance data from North American volunteers, and the results showed that the two-marker random forest models accurately identified recent infections. The AUC for the North American group was higher than the AUC for the Bangladesh group in the 200-day window, but the reverse was true for the 100-day and 45-day windows. The robust efficacy demonstrated by the cross-sectional antibody models indicates a significant potential for the dependable estimation of contemporary infection incidence via serological surveys, which may serve as an adjunct to clinical surveillance data.

In another study, sRNA was analyzed to understand how Vibrio cholerae interacts with the human genome. Small RNAs (sRNAs) play a critical role in bacterial adaptation to new environments. Fakhry *et al.* (2017) proposed a logistic regression ML approach that predicted bacterial sRNAs as belonging to one of two categories: (a) sRNAs that bind to the RNA-binding protein RsmA/CsrA in various bacterial species, and (b) sRNAs regulated by the virulence master regulator, ToxT, in *V. Cholerae.* The researchers formed a dataset consisting of 1,342 test sets of positive examples, such as RsmA-regulated and ToxT-regulated sRNAs, as well as negative examples. Sequence and structural features, such as recurrent motifs in low-energy secondary structures, SSC triplets, and stem-loop structures, were used as inputs to the logistic regression model to improve predictive accuracy. Fakhry *et al.* (2017) also developed a web interface to facilitate accessibility and promote further research. The reliability of the trained models in predicting sRNAs was tested using independent datasets. For RsmA-regulated sRNAs, 1,325 out of 1,342 (~98.7%) were correctly predicted as sRNAs. For ToxT-regulated sRNAs in V. cholerae, the model identified key features, such as the Rho-independent terminator and poly-U tail.

Vibrio cholerae was studied by Dutilh *et al.* (2014) to determine which elements of its genome reflect its structured occurrence and persistence across various locations (space), years (time of sampling), and clinical or environmental sources of the strain (habitat) worldwide. Dutilh *et al.* (2014) demonstrated the role of mobile genetic elements in shaping the pathogen's evolution. The researchers implemented a Random Forest (RF) machine learning model to classify bacterial strains by phenotype, identifying the genomic elements most relevant to the model. The analysis included 274 draft genome sequences, revealing that variations in phages, transposable elements, and plasmids contribute to the genomic diversity of *V. Cholerae* across different niche dimensions. The researchers utilized a single nucleotide polymorphism (SNP)-based phylogenomic tree to clarify the phylogenetic relationship between the VC833 *V. Cholerae* strain, isolated in Nigeria, and strains collected from Nepal in 2010, thereby suggesting a lineage associated with the recent cholera outbreaks recorded in Asia, Africa, and Haiti. The random forest (RF) model successfully categorized the genomes based on their clinical or environmental habitat with an accuracy rate of 89.2%. In contrast, the space-RF model achieved genome classification by continent with a mean accuracy of 45.3%, while the time-RF model accounted for 62.0% of the variance observed. However, the research focused on genomic analysis and did not include experimental validation of the findings, which is a limitation.

Access to good hygiene and clean water is recommended as a preventive measure against cholera epidemics. Several machine learning studies have aimed to improve access to clean water, especially in densely populated and underdeveloped regions susceptible to cholera. Nirmala (2023) applied ML algorithms with the aim of predicting water potability. The author used Random Forest and Support Vector Machine models to address the challenge of contaminated water consumption. Nirmala (2023) used a dataset from Kaggle containing 3,276 samples with nine water quality parameters: hardness, pH, chloramines, solids, organic carbon, conductivity, sulfate, turbidity, and trihalomethanes. The author employed exploratory data analysis techniques, such as the KNN Imputer to replace missing values with those of

neighboring data points, and the interquartile range (IQR) to remove outliers. The random forest model achieved 69.20% accuracy, while the support vector machine model achieved 69.05% accuracy. However, the relatively low accuracy of these models suggests that they may not be reliable for practical use, given the sensitive nature of the application.

4. The Critical Role of Clean Water in Combating Cholera

Water is a crucial and indispensable resource for sustaining human life, and maintaining its quality is of utmost importance for the well-being of individuals. Water accounts for 71% of the earth's volume, out of which only 0.3% of the water is available to us. From this a high percentage is unfitted for drinking (Arora *et al.*, 2022). According to World Health Organisations (WHO), clean water is not readily available for one out of every six people on the planet, which is around 1 billion people. Contaminated drinking water presents serious health hazards, including diseases such as cholera and other waterborne illnesses. Therefore, providing safe and clean water is essential for promoting public health (Patel *et al.*, 2023). Recent studies reveal that around 3,575,000 people die each year from cholera.

Nirmala Malagi (2023) employed machine learning techniques to predict water potability. The author implemented random forest and support vector machine models to tackle the challenge of contaminated water consumption, which could lead to contracting harmful infectious diseases such as cholera. (Nirmala Malagi, 2023) used a dataset from Kaggle that consisted of 3,276 samples with 9 water quality parameters: Hardness, pH, chloramines, solids, organic carbon, conductivity, sulfate, turbidity, and trihalomethanes. The author employed exploratory data analysis methods such as KNN Imputer, which replaces missing values with those of its neighbors, and IQR for eliminating outliers. The random forest model achieved an accuracy of 69.20%, while the support vector machine model achieved an accuracy of 69.05%. A limitation of this research is that the accuracy of the models implemented is relatively low and may not be reliable for practical use, given the sensitive nature of the use case.

Arora *et al.* (2022) implemented various machine learning techniques to forecast cholera outbreaks based on spatio-temporal and environmental data. The initial phase involved scaling the dataset and setting appropriate training and testing sets to ensure robust model training and evaluation. The dataset was carefully adjusted, with domain states set to 0.8 and later split into a 0.5 ratio for training and testing to refine prediction accuracy. During the machine learning model implementation for cholera prediction, the domain state was consistently set to 0.8 across all models. For logistic regression, a fold with 16 splits was used to handle the dataset effectively. Gradient search techniques were applied to fine-tune hyperparameters for all models. The study employed a comprehensive approach to hyperparameter tuning and model evaluation. After training, it was essential to identify the most suitable model for cholera prediction. This was achieved by comparing performance metrics derived from the confusion matrix of each model. The model achieved an accuracy of 93.58%, AUC (Area Under the Curve) score of 92.20%, and F1 score of 93.74%. This study can be improved by testing models under both extreme and normal conditions to ensure robustness across different scenarios.

5. The Transmission and Contamination of Cholera

A communicable disease spreads from person to person through contact with blood, body fluids, or airborne viruses. If not treated with appropriate care, these diseases can become life-threatening. (Karn *et al.*, 2019). Cholera is highly contagious and can be transmitted through infected faecal matter entering the mouth or via water or food contaminated with *V. Cholerae* bacteria. These organisms thrive in salty waters and can infect humans and other organisms that come into contact with or swim in the water. Cholera spreads through various means. The bacteria can survive outside the body and easily contaminate water sources and food. Additionally, individuals with the disease excrete large quantities of *Vibrio* bacteria in their stools, which can contaminate other people, as well as items like clothing, sheets, and household objects. The most common mode of transmission is through infected fecal matter entering the mouth.

Karn *et al.* (2019) implemented a medical diagnosis system using Artificial Neural Networks (ANN). The system was trained with a backpropagation algorithm and gradient optimization techniques to improve its accuracy compared to rule-based models. The ANN had 16 input nodes in the input layer and 2 hidden layers, containing 9 nodes each. The neural network had 1 output node, which predicted the presence of the disease. The dataset by (Karn *et al.*, 2019) consisted of 16 symptoms that are prominent in both cholera and dengue: joint/muscle pain, fever/chills, rashes, headache, loss of appetite, nausea, vomiting, bleeding, swollen lymph nodes, fatigue, diarrhea, abdominal pain, muscle cramps, pain behind the eyes, loss of consciousness, and change in breathing patterns. The dataset comprised 112 entries: 91 for the training set and 21 for the testing set. The ANN achieved a cholera prediction accuracy of 98.90%. This high prediction accuracy demonstrates the reliability of the diagnosis system for use by medical practitioners and patients. However, the ANN training algorithms can be improved further to achieve negligible error and higher accuracy.

6. Vibrio cholerae: A Global Pathogen.

The cholera-causing Vibrio cholerae has persisted for centuries across diverse global niches. The three niche dimensions of time, space, and environment are evident in the Vibrio cholerae genome. Alongside non-pathogenic environmental strains, Vibrio cholerae is a widely distributed pathogen that has co-evolved with humans over time (Thompson *et al.*, 2014). Dutilh *et al.* (2014) studied Vibrio cholerae, the agent responsible for cholera, to identify which elements of its genome reflect its structured occurrence and long-term persistence across different locations (space), years (time of sampling), and clinical or environmental sources (habitat) worldwide.

Dutilh *et al.* (2014) revealed the involvement of mobile functions in shaping the pathogen's evolution. The researchers employed a RF Model to categorize the bacterial strains based on phenotypic characteristics, thereby pinpointing the genomic components that hold significant relevance to the model. The study included 274 draft genome sequences and revealed that variations in phages, transposable elements, and plasmids are instrumental in shaping the genomic diversity of *V. Cholerae* across different ecological niches. Dutilh *et al.* (2014) employed an SNP-based phylogenomic tree to illustrate the relationship between the

VC833 *V. Cholerae* strain, isolated in Nigeria, and strains from Nepal in 2010, suggesting a lineage responsible for recent cholera outbreaks in Asia, Africa, and Haiti. The Random Forest (RF) model classified the genomes by clinical or environmental habitat with an accuracy of 89.2%. The space-RF model categorized the genomes by continent with an average accuracy of 45.3%, while the time-RF model accounted for 62.0% of the variation. The study centered on genomic analysis but lacked experimental validation, which is a limitation of the research.

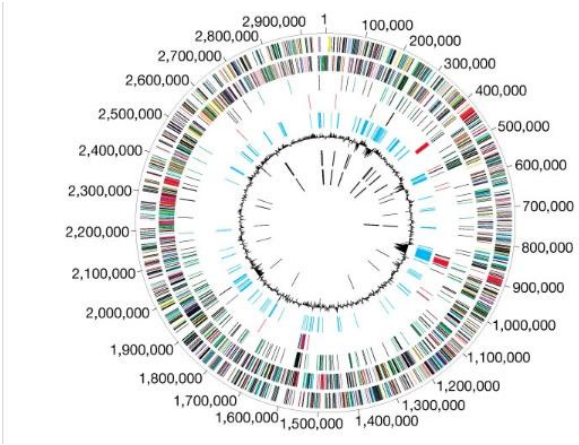

Fig. 3. Circular representation of the *V. Cholerae* genome 3

Small RNAs (sRNAs) are essential for bacteria's environmental adaption. Using a logistic regression machine learning technique, Fakhry *et al.* (2017) predicted that bacterial sRNAs would fall into one of two groups: (a) sRNAs that attach to the RNA-binding protein RsmA/CsrA in a variety of bacterial species, and (b) sRNAs controlled by Vibrio cholera's master regulator of virulence, ToxT. The scientists created a dataset with 1,342 test sets of negative and positive examples, including sRNAs regulated by RsmA and ToxT. To improve the logistic regression model's predictive accuracy, sequence and structural features—such as stem-loop structures, SSC triplets, and recurring motifs inside low-energy secondary structures—were fed into the model. Fakhry *et al.* (2017) built a web interface to facilitate accessibility and enable further research. The authors tested the reliability of the trained models in predicting sRNAs using independent test sets. For RsmA-regulated sRNAs, 1,325 out of 1,342 (~98.7%) were correctly predicted as sRNAs. For ToxT-regulated sRNAs in Vibrio cholerae, the model identified key features such as the Rho-independent terminator and poly-U tail.

Cholera incidence estimates are paradigmatically obtained from passive case-based reports of acute watery diarrhoea, and confirmed by culture, or polymerase chain react (Dick *et al.*, 2012). These estimates are limited in diagnosing the past cholera incidences in individuals, because they test for the presence of V. cholera O1 in human samples. Additionally, the process of collecting and managing incidence reports has been plagued by problems such as the imbalance data and improper data handling. To improve the quality of cholera incidence estimates, Azman *et al.* (2019) created ML models that uses 6 serological markers. The data population consists of confirmed cholera patients and their close contacts, obtained from Bangladesh. The emergence of vibriocidal antibodies, along with the immune responses specific to the *V. Cholerae* O1 serogroup antigens, which feature the O antigen of lipopolysaccharide and the cholera toxin B subunit, has

evidenced effectiveness as the most reliable immunological indicators for recent *V. Cholerae* O1 infection. If using cross-sectional proves effective, it will provide an effective alternative for evidence-based approach for targeting and tracking cholera interventions. The training data for the model was collected from 2006 to 2015 from the patients of cholera and their household contact in Dhaka Bangladesh, between the ages of 2 to 60). The serology data from the infected subjects was extracted from the blood samples of the subjects on 7, 30, 90, 180, 270, 365, 540, 720, and 900 days (about 2 and a half years) after symptom. The blood samples from the household contact were collected with respect to the day of enrolment of the index case on days: 2, 7 and 30. Alongside blood samples, the rectal swabs were collected daily during the first 10 days (about 1 and a half weeks) of enrolment. The data from the uninfected household contacts served as the controls in analyses and was used as the background antibody distribution. To build the model, the single-marker thresholds were evaluated by performing 20-fold cross-validated area under curve (cvAUC), assigning an individual from the infected training data into a fold. The exact titer threshold that led to the maximized specificity and sensitivity was also determined using a single observation from a single observation. Afterwards, the random forest model was used to identify recently infected individuals base on cross-sectional antibody titers. An individual is considered infected over a period if they had a cholera incidence over the same period. The period of considerations starts with 1 as the lower bound, to the higher bound, selected from the following values: 10, 45, 100, 200, or 365 days (about 12 months). Hence, 45 would mean being infected from day 1 to day 45. The random forest classification consisted of 1000 trees. Model performance was evaluated using 20-fold cvAUC. The external validation of the model was carried using the 6 months Sero surveillance data of north American volunteers. The result showed that two-marker random forest models accurately identified recent infection. The AUC for the north American group was higher than the AUC of the Bangladesh in the 200-day window. Reverse being the case for 100-day window and the 45-day window. From the limited antibodies from the north American group suggests the potency of this approach. The effectiveness of the cross-sectional antibody models raises the prospect of accurately calculating the incidence of recent infections using serological surveys, which might serve as an adjunct to the utilization of clinical monitoring data.

7. CONCLUSIONS AND RECOMMENDATIONS

Cholera is a diarrheal disease caused by vibrio cholera and is responsible for the spread of other diseases leading to illnesses, and even death. As evidenced by statistics, it is an alarming issue with a devastating effect on human health, especially in developing countries. These countries are usually susceptible to cholera outbreaks due to poor sanitation hygiene, improper or damaged drainage systems, natural or man-made disasters, and neglect of good health policies. Traditional and mechanical statistical method for predicting cholera have been employed to mitigate the outbreak of cholera but are limited given their static properties. With the advent of Artificial Intelligence, researchers have utilized a more dynamic and improved approach—Machine Learning Algorithms—to mitigate the spread of this treacherous disease.

Given the proliferation of this disease in developing countries, we have successfully reviewed how relevant studies have employed ML algorithms to curb further cholera outbreaks Some studies focused on building

Machine Learning models to forecast the outbreak of cholera. They took datasets from previous cholera-recorded cases, alongside other variables such as socio-economic factors, environmental and social variables, Environmental Climatic Variables (ECVs), and climatic variables. PCA, ADASYN, and SMOTE rectified data imbalance and missing values. These data were then trained with different machine learning algorithm models: RFC, Native Bayesian Classification, ANN and MLP, XGBoost and others. Subsequently, these studies discovered suitable variables strongly correlated with accurately predicting cholera outbreaks.

Some studies also highlighted the use of ML techniques to understand the interaction of Vibrio cholera with humans—both genomic and cellular level—and its interaction with water. Findings revealed hidden strains in the genomic elements and other dependent variables that showed a significant correlation with cholera. These studies utilized Machine Learning algorithms—some studies combined both mechanistic approaches with ML—to (a) prevent the outspread of cholera by revealing variables that are consistent with Vibrio cholera, (b) show variables that are strong predictors for cholera outbreaks, and (c) predict the outbreak of cholera in different regions, especially in developing countries. Given the noxious effect of cholera and its epidemic tendencies, more research using ML needs to be carried out and built upon previous relevant works, as the existing ones are relatively small. Furthermore, future studies can take into account socio-economic variables since this was a limitation for some previous studies, as they took into preference climatic variables. Socio-economic factors illumine potential hidden variables, different from climatic variables, that might contribute to a more accurate prediction of cholera outbreaks. Future studies should also factor in lag time when measuring dependent variables, such as rainfall, for more accurate prediction. Lastly, future works can make their model more generalizable to overcome only endemic cholera outbreak cases. These works have revealed that predicting cholera outbreaks will significantly aid the government and disease regulatory bodies in taking action and implementing policies to help mitigate the widespread spread of this disease. This will alleviate the rapid increase of cholera in developing countries and improve the standard of living in the society at large.

REFRENCES

[1]. Ahmad Amshi, H., Prasad, R., Sharma, B. K., Yusuf, S. I., & Sani, Z. (2024). How can machine learning predict cholera: insights from experiments and design science for action research. *Journal of Water and Health*, *22*(1), 21–35. https://doi.org/10.2166/wh.2023.026

[2]. Alfred, R., & Obit, J. H. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon*, *7*(6), e07371. https://doi.org/10.1016/j.heliyon.2021.e07371

[3]. Arora, A., Singh, V., Gourisaria, M. K., & Jena, A. K. (2022). Analyzing the Potability of Water using Machine Learning Algorithm. *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 250–256. https://doi.org/10.1109/CCiCT56684.2022.00054

[4]. 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

[5]. Ashari, A., Paryudi, I., & Min, A. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications*, *4*(11). https://doi.org/10.14569/IJACSA.2013.041105

[6]. Azman, A. S., Lessler, J., Luquero, F. J., Bhuiyan, T. R., Khan, A. I., Chowdhury, F., Kabir, A., Gurwith, M., Weil, A. A., Harris, J. B., Calderwood, S. B., Ryan, E. T., Qadri, F., & Leung, D. T. (2019). Estimating cholera incidence with cross-sectional serology. *Science Translational Medicine*, *11*(480), eaau6242. https://doi.org/10.1126/scitranslmed.aau6242

[7]. Campbell, A. M., Racault, M.-F., Goult, S., & Laurenson, A. (2020). Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables. *International Journal of Environmental Research and Public Health*, *17*(24), 9378. https://doi.org/10.3390/ijerph17249378

[8]. Charnley, G. E. C., Yennan, S., Ochu, C., Kelman, I., Gaythorpe, K. A. M., & Murray, K. A. (2022). The impact of social and environmental extremes on cholera time varying reproduction number in Nigeria. *PLOS Global Public Health*, *2*(12), e0000869. https://doi.org/10.1371/journal.pgph.0000869

[9]. Dick, M. H., Guillerm, M., Moussy, F., & Chaignat, C.-L. (2012). Review of Two Decades of Cholera Diagnostics – How Far Have We Really Come? *PLoS Neglected Tropical Diseases*, *6*(10), e1845. https://doi.org/10.1371/journal.pntd.0001845

[10]. Dutilh, B. E., Thompson, C. C., Vicente, A. C., Marin, M. A., Lee, C., Silva, G. G., Schmieder, R., Andrade, B. G., Chimetto, L., Cuevas, D., Garza, D. R., Okeke, I. N., Aboderin, A. O., Spangler, J., Ross, T., Dinsdale, E. A., Thompson, F. L., Harkins, T. T., & Edwards, R. A. (2014). Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. *BMC Genomics*, *15*(1), 654. https://doi.org/10.1186/1471-2164-15-654

[11]. Fakhry, C. T., Kulkarni, P., Chen, P., Kulkarni, R., & Zarringhalam, K. (2017). Prediction of bacterial small RNAs in the RsmA (CsrA) and ToxT pathways: a machine learning approach. *BMC Genomics*, *18*(1), 645. https://doi.org/10.1186/s12864-017-4057-z

[12]. Fung, I. C.-H. (2014). Cholera transmission dynamic models for public health practitioners. *Emerging Themes in Epidemiology*, *11*(1), 1. https://doi.org/10.1186/1742-7622-11-1

[13]. Karn, S., Sangole, S., Gawde, A., & Joshi, J. (2019). Prediction and classification of vector-borne and communicable diseases through artificial neural networks. *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019, Iciccs*, 1011–1015. https://doi.org/10.1109/ICCS45141.2019.9065500

[14]. Lawal, L., Amosu, O. P., Lawal, A. O., Wada, Y. H., Abdulkareem, A. O., Shuaib, A. K., Jaji, T. A., Mogaji, A. B., Abdul-Rahman, T., Adeoti, S. G., & Buhari, A. O. (2024). The surging cholera

epidemic in Africa: a review of the current epidemiology, challenges and strategies for control. *International Journal of Surgery: Global Health*, *7*(2). https://doi.org/10.1097/GH9.0000000000000440

[15]. Leo, J., Luhanga, E., & Michael, K. (2019). Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. *The Scientific World Journal*, *2019*, 1–12. https://doi.org/10.1155/2019/9397578

[16]. Nirmala Malagi. (2023). Water Potability Prediction using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*. https://doi.org/10.56726/IRJMETS44413

[17]. Nuhu, Y. (2021). CHOLERA PREDICTION MODEL USING FEATURE CLUSTERING BAYESIAN TECHNIQUE. *Journal of Applied Science, Information and Computing*, *2*(2), 18–23. https://doi.org/10.59568/JASIC-2021-2-2-02

[18]. Nusrat, F., 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), 48. https://doi.org/10.3390/cli10040048

[19]. Ogore, M. M., Nkurikiyeyezu, K., & Nsenga, J. (2021). Offline Prediction of Cholera in Rural Communal Tap Waters Using Edge AI inference. *2021 IEEE Globecom Workshops (GC Wkshps)*, 1–6. https://doi.org/10.1109/GCWkshps52748.2021.9682128

[20]. Onyijen, O. H., Olaitan, E. O., Olayinka, & Oyelola. (2023). DATA-DRIVEN MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF CHOLERA OUTBREAK IN WEST AFRICA. In *Western European Journal of Modern Experiments and Scientific Methods* (Vol. 1, Issue 1). https://westerneuropeanstudies.com/index.php/1https://westerneuropeanstudies.com/index.php/1

Cite this Article:

Jessica Nwobodo, Shugaba Wuta, Michael Ibitoye, Paul Omagbemi, Martins Offie*, "****Recent Advances in Machine-Learning Driven Cholera Research: A Review****" International Journal of Scientific Research in Modern Science and Technology (IJSRMST), ISSN: 2583-7605 (Online), Volume 3, Issue 10, pp. 07-21, October 2024.*

Journal URL: https://ijsrmst.com/

DOI: https://doi.org/10.59828/ijsrmst.v3i10.255.