

Modeling Cholera Outbreaks in Kenya using Machine Learning Algorithm

Paul Njuguna
Kenyatta University

Keywords—cholera outbreaks, machine learning, Random Forest, SMOTE, Kenya, early warning systems.1.

I. INTRODUCTION

1.1 Background

Cholera is a diarrheal disease caused by the bacterium, *Vibrio cholerae*; the primary mode of transmission is through contaminated food and water. There are two main serogroups O1 and O139 that are known as the most responsible for major epidemics of cholera worldwide (Dossou Sodjinou et al., 2022). Cholera is highly widespread in those regions with ineffective sanitation and low availability of safe drinking water, which disproportionally concerns developing nations, such as Kenya World Health Organization (WHO, 2023).

Vibrio cholerae enters via contaminated water or food and is able to survive the acidic stomach environment, colonizes the small intestine with colonization factors and is able to proliferate in those cells. After colonization, the bacteria express a cholera toxin that is an enterotoxin of the AB₅ type, which activates adenylate cyclase to increase intracellular cyclic AMP, causing profuse electrolyte and fluid secretion. The clinical signs include excessive watery diarrhoea (typically 'rice-water stool'), vomiting and muscle cramps, which, if untreated, can cause rapid dehydration and be fatal (Frontiers in Microbiology, 2023).

Meteorological information shows that cholera is a continuous menace in Kenya and has been reported to occur regularly during rainy seasons when flooding makes water sources contaminated. In Kenya, for instance, public health measures have had no impact on the continued vulnerability, as the Ministry of Health has reported on recurrent outbreaks in counties (Kenya Ministry of Health, 2023). Moreover, the transmission risk of cholera is much increased in informal settlements due to overcrowding, insufficient basic water and sanitation services, and urbanization. In informal settlements, the rainy seasons are important times for the spread of cholera because of the increased risk of flooding and contamination of water sources. In recent instances, for example, outbreaks have been reported in seven counties including Nairobi, Mombasa, Kisumu, Migori, Homa Bay, Kwale and Turkana, leading to many deaths in the wake of heavy rains. (Daily Nation 2025, June 18)

Stool culture and rapid diagnostic tests (RDTs) are essential tools for early detection of the cholera. However, access to these services is not as widespread in rural areas as in urban areas, and that slows down reports and reduces response strategies. Logistics problems, the low awareness of the public about cholera prevention and control measures (WHO 2023), are among the challenges faced by treatments such as oral rehydration solutions (ORS) and antibiotics in situations of resource limitations.

1.2 Problem Statement

Although cholera has been and remains a serious public health problem in Kenya, current prediction models rely primarily on past epidemiology data and simple statistical methods. The models do not usually consider the complex interactions between environmental factors and socio-economic factors that affect the outbreak of cholera. Furthermore, while there is some research on cholera prediction using machine learning in Kenya, existing studies are limited and do not provide a thorough relative analysis of numerous machine learning algorithms. This study aims to overcome such gap by developing a comprehensive and effective model that incorporates both socio-economic and environmental indicators to predict cholera. The purpose of this study is to assess and compare various machine learning algorithms to see which one is most effective at predicting cholera outbreaks in Kenya, thereby potentially providing better early warning systems and public health response capabilities.

To evaluate the effectiveness of the Random Forest machine learning model in predicting cholera outbreaks in Kenya.

1.3.2 Specific Objectives:

Examine key environmental and socio-economic factors that influence the outbreak of cholera in Kenya.

To create a trained Random Forest model and assess the model's performance on standard metrics for cholera outbreak prediction.

To evaluate the effect of handling techniques for class imbalance (e.g. SMOTE) on model performance.

This study is important as the researchers will have a comparative study of the various machine learning techniques to be used to ensure that the best model will be used to serve the predicted need for dealing with an outbreak in Kenya. This study will use environmental and socio-economic information to help inform early warning systems, which will improve cholera response strategies. Additionally, the study will test if multiple algorithms are applicable and able to be used in cholera outbreak prediction, which may further enhance prediction accuracy.

This work is hoped to contribute to the understanding of how cholera works in Kenya to more effectively integrate it into public health strategies. Cholera is a critical public health priority and outbreak threat, especially in resource-scarce countries, generally endemic to this infectious disease. The machine learning approach towards outbreaks proposed here aims to address a significant gap in the existing knowledge and will offer useful inputs to the health authorities.

II. LITERATURE REVIEW

This study is informed by four theories that are woven together to offer a solid basis for understanding and modelling a cholera outbreak through machine learning. These include:

This is the main theoretical framework underpinning this study. It defines the cause of disease, in terms of interaction among three factors: *Vibrio cholerae*, human population, and environment (sanitation, water quality, etc.) (Park, 2009).

This theory directly guides the identification and incorporation of the epidemiological, environmental, and socio-economic factors into the machine learning model which is realized in the study. It is used to guide the selection of

variables that can be used to predict outbreaks of cholera and provides the conceptual foundation for the modeling approach.

This theory is characterized by the interconnections of parts in an overall system. Diseases are not driven by single factors in public health, but through factors interacting with each other, e.g., urban infrastructure, climate, behaviour. (von Bertalanffy, 1968).

It is justification for using machine learning, a technique that can be used to analyze complex interdependent variables for predicting cholera outbreaks, in this case, Random Forest. Systems Theory helps us consider the prediction task as a systems-level challenge.

HBM describes the relationship between people's beliefs and their health issues, and how they think about the risks and barriers they face that affect their behavior and response. (Rosenstock, 1974).

This is a model that facilitates the qualitative aspect of this research. Data collected through the household survey and interviews, including hygiene practices and water consumption, will aid in the interpretation of the machine learning predictions and human behavior that contributes to the risk of disease.

This theory concentrates on the acquisition, processing and decision-making of information. (Atkinson & Shiffrin, 1968).

It gives a theoretical basis for the machine learning model to process the input data such as rainfall, sanitation level, and population density to generate the prediction of output such as disease cases which inform public health decision-making.

To date, cholera remains a significant health problem in sub-Saharan African countries including Kenya, and is known to occur seasonally, linked to sanitation conditions and floods. Previous

epidemics (1997, 2009) led to thousands of cases and hundreds of deaths (Santangelo et al., 2023). Despite interventions like oral cholera vaccines, WASH programmes, cholera remains a persistent problem particularly in informal settlements and underserved rural areas.

In the context of public health, predictive modeling is a valuable tool that can help predict disease outbreaks and inform proactive interventions. Traditionally, statistical models based on historical data and linear assumptions were used. But, these models are not necessarily flexible enough to account for and to analyze complex interactions among multiple variables. Examples of such models

include one developed by (Kavinya et al. 2023) for cholera prediction in Mozambique where climate variables were used, but where limitations in accuracy were attributed to the lack of socio-economic data.

Machine learning (ML), especially ensemble models such as Random Forest (RF), is growing to become a promising approach to model disease outbreaks. ML techniques can detect intricate and non-linear patterns in large data sets. Better performance of the RF model was found in predicting cholera outbreaks using climate and socio-economic data compared to traditional models (Dossou Sodjinou et al. 2022a).

Likewise, (Das et al. 2023) used RF, decision trees, and support vector machines for the analysis of cholera cases in Bangladesh. The importance of using multiple types of variables – climate, infrastructure, and access to healthcare – in prediction models was shown by the superiority of RF over other algorithms.

Outbreaks of cholera are strongly linked to environmental conditions like rainfall, temperature and water quality. The flood is often a consequence of high rainfall and it may lead to contamination of water sources (Legros, 2018). The temperature also affects the growth of bacteria, and higher temperatures favor the growth of *Vibrio cholerae*.

Socio-economic factors like population density, poverty, education and sanitation are critical factors. Rapid urbanization and informal settlements have put people at risk of cholera infection, as there is insufficient infrastructure and public health resources (Santangelo et al., 2023).

The main features extracted for the predictive models in this research stem from these environmental and socio-economic factors. The study will gather and analyze historical information about rainfall, temperature, water quality, population density and access to sanitation to train machine learning algorithms, specifically Random Forest (RF), to identify patterns for cholera outbreaks. These variables will be preprocessed and entered into the model for testing their prediction ability and finding out the most important variables that influence the occurrences of outbreaks in various regions of Kenya.

Many machine learning algorithms have been used for disease prediction problems in various health applications, including infectious diseases like cholera, malaria and COVID-19. Popular models are Decision Trees, Random Forest, Support Vector Machines, Neural Networks, and Gradient Boosting techniques such as Gradient Boosting Machines and XGBoost, with their strengths depending on the dataset and prediction goals (Katuwal & Suganthan, 2021; Shailaja et al., 2022). They have been shown to perform well on the large scale of epidemiological and environmental data, and they can be useful tools for early warning system development and public health planning.

Recent studies have further validated the use of RF in cholera prediction. Tshimula et al. (2024) created an RF model that integrated the climate, sanitation, and population data to forecast cholera outbreaks in Kenya. The model proved to be accurate and showed the importance of real-time and multi-factorial data.

Ergüzen and Ünver (2018) also showed RF's ability to determine the most influential factors that lead to outbreaks. This makes RF not just a predictive tool but a diagnostic one, which can help public health officials prioritise interventions.

Although these developments have been made, there are still some issues in the prediction of cholera. Availability and quality of data continues to be a significant constraint, especially in low-resource countries, like Kenya. The accuracy and timeliness of predictive models are significantly impacted by problems like incomplete health records, underreporting and delays in data collection (Weppelmann et al., 2022).

In addition, factors like migration, political unrest and alterations in health-seeking behaviors play a dynamic role in the spread of cholera and are difficult to quantify. These factors make it more difficult to create models that are universally accurate.

While there have been some positive results in predicting cholera outbreaks using machine learning in certain contexts, most of the more prominent ones are from countries with a long history of cholera surveillance and highly detailed environmental data. Recently, for instance, in Bangladesh, the authors have applied Random Forest and related techniques to map the risk of cholera under current and future climate conditions and to forecast seropositivity in the national sero-surveys. Similar climate–socioeconomic modeling and spatial, temporal analyses have been reported for Mozambique. To the contrary, there is little published work that uses machine-learning outbreak-prediction techniques specifically for Kenyan cholera data, or that incorporates fine-grained socio-economic and informal-settlement data into predictive models. The geographic imbalances suggest a need: There is a need for more Kenya-specific

ML work, particularly integrating epidemiological surveillance, environmental data, with household-level behavioural indicators from informal settlements.

This study aims to fill these gaps by creating a Random Forest model specific to the Kenyan context, which utilizes environmental, epidemiological and socio-economic data to provide a more accurate assessment of cholera outbreaks.

III. METHODOLOGY

For this study, the researcher adopted a mixed methods research design, which combines quantitative and qualitative methods, to give a complete picture of the cholera outbreak in Kenya. The quantitative component involves analyzing numerical data, including historical cholera case records, meteorological data (e.g. temperature, rainfall) and socioeconomic indicators (e.g. population density, sanitation, income). This data will be used to develop, train and test machine learning models for predicting cholera outbreaks.

Key informant interviews (KIIs), focus group discussions (FGDs), and household surveys make up the qualitative component, which aims to investigate community behaviors, hygiene practices, and perceptions regarding cholera transmission. This qualitative data, along with its contexts and social understandings, adds richness and complement to the quantitative findings. The statistical trend analysis and community-level realities are integrated in the study, thus ensuring a basis for predictive modelling and a comprehensive interpretation and application of the results.

The study will be conducted in specific areas of Kenya vulnerable to cholera outbreaks and have had frequent outbreaks in the past. These include counties where there are high cholera incidences such as Nairobi, Kisumu, Migori and Kwale. Some of the features of these areas are high population density, low water quality and sanitation facilities and susceptibility to flooding during rainy

seasons. The study areas were selected because of environmental susceptibility, historical outbreak information and socio-economic factors which increase the risk of cholera transmission.

The primary and secondary data collection technique will be used to gather data for this research study from multiple sources. Data types, sources, and data collection techniques are described as follows:

Environmental Data

Data Types: Temperature, Rainfall, Frequency of flooding, and Water Quality indicators.

The data has been compiled from Kenya National Bureau of Statistics (KNBS) and Kenya Demographic and Health Surveys (KDHS).

Collection Method/Tools:

Epidemiological Data

Data Type: Historical data on the location, timing, and numbers of cholera outbreaks and deaths.

Sources: Ministry of Health (MoH), World Health Organization (WHO).

Collection Method/Tools:

Obtain outbreak data on location and cases from MoH Integrated Disease Surveillance and Response (IDSR) weekly bulletin.

Download WHO situation reports on cholera from the WHO Africa Region website.

Use the Spreadsheet (e.g., Excel) and SPSS for preliminary descriptive analysis and data cleaning.

Socio-Economic Data

Type of Data: Population density, access to clean water and sanitation, household income levels.

Data from the Kenya Demographic and Health Surveys (KDHS) and the Kenya National Bureau of Statistics (KNBS).

Collection Method/Tools:

Access KDHS datasets with permission at the DHS Program Data Portal.

Extract Population and Household level information from KNBS databases.

Collect and manipulate data with SQL DBMS for easy retrieval and integration with epidemiological data.

Human Behavior Data

Data type: Community sanitation practices, water storage and usage behaviour, hygiene practices.

Sources: Local communities in the selected study areas.

Collection Method/Tools:

Carry out structured household surveys with KoboToolbox or ODK Collect (mapping data collection apps).

Conduct focus group discussions (FGDs) with community members (with consent and using digital audio recorders).

Use semi-structured interview guides to carry out Key Informant Interviews (KIIs) with local health workers.

Extract qualitative data and code and analyse them thematically with NVivo.

In total, 500 households will be involved in the study. This sample size not only contains sufficient data for building and validating machine learning models, but also ensures that the socioeconomic and environmental spectrum from the parts of Kenya that experience cholera is represented.

We'll use a multistage stratified random sampling method. The sub-counties and wards will be selected randomly after strata of the counties based on the cholera risk. Then, a systematic sample of households in the areas will be selected. This approach ensures that there is representation across all urban, peri-urban, rural and informal settlement populations.

Households will be recruited with the support of local community administrative leaders and Community Health Volunteers (CHVs). Community sensitization will be done before data collection.

The primary respondent will be an adult household head (age 18 and above) or other adult household member with knowledge of the water, sanitation and health practices in the household.

Samples will be targeted towards low-income households, flood-prone areas and informal settlements, to ensure that vulnerable populations are represented. Participants will be invited and not forced to participate and gender parity will be encouraged.

The third step is on qualitative data analysis framework. The third step is 3.6 Qualitative Data Analysis Framework. Qualitative data will be analysed using Braun and Clarke's six-step framework of familiarisation with data, coding, theme development, theme review, theme definition and reporting. NVivo software will be used to assist with data management and analysis.

The gathered data will be modeled via the following steps:

- Data Cleaning
- Normalization
- Feature Selection
- Categorical Encoding
- Data Splitting

Different machine learning algorithms are able to make different kinds of predictions on a disease. Different algorithms will be compared to assess their efficiency, accuracy, and applicability in predicting cholera outbreaks. These algorithms have been chosen are:

Decision Trees: Recognized for their interpretability and ease of implementation, the Trees are useful for classification problems but can suffer from overfitting when dealing with complex datasets.

Support Vector Machines (SVM): This is a powerful classifier which gives high accuracy but it is expensive and less suitable for real time prediction.

Neural Networks: These models are found to be useful in pattern recognition and can give a very high accuracy in predicting the disease. They must be trained with a lot of data and computational power, however.

Random Forest: An ensemble learning framework which involves constructing numerous decision trees and combining them to enhance the accuracy of the predictions as well as avoid overfitting. Random Forest has been known to perform well in predicting diseases in its robustness and capacity to analyze large-volumes-of-data.

Ensemble Learning (Stacking Models): Ensemble methods provide better prediction reliability and accuracy by integrating multiple models. A hybrid approach will be investigated to see whether or not it is superior to individual algorithms.

The historical data of cholera outbreaks will be used to test each algorithm, and both the algorithms and the final performance will be evaluated according to previously established metrics. The results will help know the most efficient predictive model for cholera outbreaks in Kenya.

Random forest is one of the ensembles forms of learning that create multiple decision trees and formulates their results creating an overall guess for the final forecast. It does this by training random subsets of data and random selections of features for each tree to make the model strong and not too sensitive to any single feature.

The Random Forest algorithm will be used in this study as follows:

Random Forest Training:

Model Initialization: The model will be built using the Random Forest algorithm using Python Scikit-learn library. This model constitutes a combination of decision trees between which individual trees are formulated on a random collection of the training data.

The major hyperparameters, denoting number of trees `n_estimators`, `max_depth` of trees and number of samples per leaf `min_samples_leaf`, will be optimized for performance of the hyperbat. Grid search and Cross validation technique will be used to find the best combination of hyper parameters.

Feature Selection: Random Forest automatically selects features based on the predictive power of each feature. This mechanism of feature importance will inform the identification of the most important features in the prediction of cholera outbreaks (e.g. rainfall, temperature, sanitation conditions, and population density).

Model Training Process:

The training dataset will be used to train the Random Forest model which will then learn a collection of decision rules based on randomly sampled features from the training set.

The data is split on the most important attribute of each tree during its training and creates various branches. This produces diversity between trees as well as limits overfitting. A tree could split in half because of a different amount of rainfall, for example, or because of a different density of population.

The Random Forest model is then created by taking the average of the predictions of all the decision trees, so that it can learn patterns of the data.

Model Testing and Evaluation:

Once trained, the model will be evaluated with the testing data set containing data that was not used during training. This will appraise the model's capability to oversimplify to new, concealed data.

Measures such as performance will be used to gauge performance:

Accuracy: It is the percentage of correct predictions that the model produces.

Precision and Recall: These will help to assess the accuracy of the model in detecting cholera epidemics (true positives) and in cases where the model detects an epidemic, but the case is absent (false positives).

F1-Score: This is the harmonic mean of precision and recall in which the computed value indicates the harmonic mean of the precision and the recall providing single value to measure the model performance.

Area Under the ROC Curve (AUC): This will determine how well the model will be able to tell the difference between cholera outbreak and non-outbreak.

iii. Confusion Matrix: A confusion matrix will show which of the predictions are true positive, true negative, false positive and false negative predictions of the model. This will give information on what kind of errors model is making.

Feature Importance Evaluation:

Random Forest provides a way to check the relevance of each feature (rainfall, temperature, socio-economic information etc.). This can be accomplished through feature_ importances_ attribute in Scikit-learn, which returns a score for each feature, representing the importance compared to the extrapolative power of the model.

The importance of the features will be determined through analysis and this will inform public health interventions and policies.

A variety of assessment metrics will be considered to see how the different machine learning algorithms perform:

Accuracy: measures the percentage of output that is outputted correctly by predicting outbreaks.

Precision and Recall: Precision will be the percentage of the correctly predicted cholera outbreaks divided by the number of cholera outbreaks predicted while Recall will be the ability of the model to identify the cholera outbreaks.

Cross-validation: A K-fold cross-validation strategy will be applied in order to make sure that the model is generalizable with new points. The approach includes dividing the data into many training and testing sets.

Hyperparameter Tuning: Random Search and Grid Search will be used in hyperparameter tuning to maximize the parameters which will lead to better performance.

All these measures of evaluation will provide a detailed assessment of the performance of each of the algorithms provided, so that the best algorithm can be selected for predicting cholera outbreaks.

After training the model, a sensitivity analysis will be performed to assess the impact of various features on the model's predictions. This will help to identify the factors that most affect cholera outbreaks – for example, rainfall, temperature and socio-economic factors. The study will use feature

importance analysis and sensitivity testing to gain meaningful insights into the factors that can cause cholera outbreaks, and to determine where interventions would be most effective.

The following procedures will be established to deal with psychological distress situations:

Some study participants may experience discomfort on an emotional or psychological basis, although these effects are uncommon with the study. The ability to identify distress will be taught to data collectors. If participation stops or is suspended, affected persons will be directed to the nearest public health facility, community health officer or county mental health focal person.

All participants will give informed consent. Data will be stored on password-protected computers which are encrypted and accessible only to the research team. Data will be safely destroyed after being kept for five (5) years. Ethical oversight of the study will be provided by the Kenyatta University Ethics Review Committee (KUERC) which will give ethical approval prior to data collection.

These ethical guidelines will guarantee responsible handling of data and the protection of participants' rights in the study.

IV. RESULTS

This chapter outlines the outcomes of implementing and testing machine learning models used to forecast cholera outbreaks in Kenya. The analysis is based on the analysis methodology discussed in Chapter Three and the study objectives discussed in Chapter One.

The results are displayed visually in plots, confusion matrices, and feature importance plots of the class distribution, confusion matrix, and feature importance.

This data comprised of epidemiological, environmental and socio-economic variables relevant to cholera transmission. These included:

- Weekly cholera case counts
- Lagged case variables
- Rolling case averages
- Rainfall
- Temperature
- Sanitation coverage indicators
- Population-related variables

Imputation methods were used for the missing values.

Categorical variables were coded into numerical form as was necessary.

Lag features were used to account for momentum effects of outbreaks from temporal variables...

To assess the generalization performance, the data set was randomly split into 80% training data and 20% testing data. This is a way to make sure that models perform as they would in real life prediction, not as they memorized it.

The findings indicate that temperature, rainfall, and floods have moderate variability over the years, respectively. The average water quality (WQI) is 60.09 indicating moderate water quality. The availability of safe water and sanitation have been improving at a slow rate with fairly low variability, indicating a continuous improvement over time.

Notes:

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

The correlation between year and temperature is very strong and positive which implies that there is a steady rise in the temperature over the years. The Water Quality Index (WQI) is highly negatively associated with temperature and floods, implying that the increase in temperatures and floods considerably worsen the water quality. There exists a moderate positive correlation between rainfall and floods, which supports the contribution of rainfall to floods.

a. Predictors: (Constant), Temperature, Rainfall, Flood Events

According to the model, 99.6% of the variance in the water quality is explained by the model which is very strong. The adjusted R² (0.994) demonstrates that the predictors are effective in explaining variations in WQI. The value of Durbin-Watson (~2.1) indicates that there is no issue of autocorrelation in the residuals.

Dependent Variable: WQI

Predictors: Temperature, Rainfall, Flood Events

According to the results of the ANOVA, the regression model is statistically significant ($p = 0.001$). This shows that a combination of temperature, rainfall and flood occurrence has a huge impact on the quality of water.

a. Dependent Variable: WQI

The most negative influence on water quality is temperature ($\beta = -0.712$), and thus, an increase in temperature has a severe impact on WQI. There is also a significant negative effect caused by flood events, and a smaller yet significant effect caused by rainfall. All the predictors are statistically significant ($p < 0.05$), which proves them to have influence on the water quality. The values of VIF are less than 5 which means that there are no severe problems of multicollinearity.

The values of the residual are small in value and the values are normally distributed around a value of zero, which means that the regression model is a good fit to the data. No extreme outliers are present, which implies that the model predictions can be trusted.

Theme 1: Water Accessibility, Quality, and Treatment Practices

Access and quality of water became a decisive factor of risk of cholera in all participants. The respondents indicated that they were dependent on different water sources, which included piped water, boreholes, lakes, shallow wells, and vendors. Nevertheless, such sources were untrustworthy or polluted frequently.

Participants from rural and lake regions highlighted direct exposure to unsafe water:

“We mostly get water from the lake... the quality is not good.” (P8)

Similarly, peri-urban and informal settlement residents expressed concerns about inconsistent supply and questionable quality:

“We buy water from vendors... sometimes the quality is not guaranteed.” (P1)

A significant finding was the inconsistent treatment of drinking water, largely influenced by economic constraints and perceived water safety:

“If we have fuel, we boil... if not, we just drink it.” (P1) “We assume piped water is safe, so we don’t treat it.” (P9)

This indicates that perceived safety and affordability strongly influence water treatment behavior, increasing vulnerability to cholera.

Theme 2: Inadequate Sanitation Infrastructure

Sanitation challenges were widespread, particularly in informal settlements and rural areas. Shared sanitation facilities, poor maintenance, and lack of infrastructure were commonly reported.

Participants described overcrowded and poorly managed facilities:

“Up to ten families share one toilet... it gets dirty quickly.” (P7)

In rural and ASAL areas, the situation was more severe, with open defecation reported:

“We don’t have toilets... most people use open areas.” (P4)

Flooding further exacerbated sanitation issues by spreading waste into the environment:

“When it floods, waste mixes with water sources.” (P5)

These findings highlight that infrastructure limitations significantly contribute to environmental contamination and disease transmission.

Theme 3: Hygiene Practices and Behavioral Inconsistency

Although most participants demonstrated awareness of basic hygiene practices such as handwashing, implementation was inconsistent. Hygiene behavior was found to be influenced by availability of resources, habits, and situational factors.

“We try to wash hands... but sometimes there is no soap.” (P1) “Children don’t always follow hygiene rules.” (P2)

A key pattern identified was reactive hygiene behavior, where practices improve only during outbreaks:

“People become careful during outbreaks... but later go back to normal.” (P3)

This suggests that behavioral inconsistency is a major barrier to effective cholera prevention, despite awareness.

Theme 4: Socio-Economic Constraints and Poverty

Economic factors emerged as a central theme influencing water, sanitation, and hygiene practices. Participants frequently cited cost as a barrier to maintaining safe practices.

“Fuel is expensive... we cannot always boil water.” (P1) “We cannot afford treatment methods regularly.” (P2)

In resource-limited settings, households prioritized immediate needs over preventive health behaviors:

“We focus on getting enough water, not treating it.” (P4)

This demonstrates that poverty directly limits the ability to adopt preventive measures, thereby increasing cholera vulnerability.

Theme 5: Environmental and Climatic Factors

Environmental conditions such as flooding, poor drainage, and water scarcity were identified as major contributors to cholera outbreaks.

Flooding was particularly significant in spreading contamination:

“Flooding contaminates wells and water sources.” (P5)

In urban informal settlements, poor drainage systems worsened the situation:

“Drainage is blocked... water stagnates and becomes dangerous.” (P7)

In ASAL regions, water scarcity limited hygiene practices:

“We cannot wash hands regularly because water is limited.” (P4)

These findings confirm that environmental factors interact with socio-economic conditions to increase cholera risk.

Theme 6: Cultural Beliefs and Perceptions

Cultural beliefs and perceptions were found to influence health behaviors, particularly in rural areas.

“Some believe lake water is natural and safe.” (P8)

Additionally, misconceptions about disease causation affected preventive practices:

“Some people don’t believe hygiene is the main cause.” (P8)

This highlights that cultural beliefs can either support or hinder effective disease prevention strategies.

Theme 7: Awareness Versus Practice Gap

While awareness of cholera and its prevention was generally high, there was a clear gap between knowledge and actual behavior.

“People know what to do, but they don’t do it consistently.” (P10)

Participants emphasized that awareness alone is insufficient without enabling conditions:

“Awareness helps, but without resources, it is difficult to act.” (P5)

This theme underscores the importance of bridging the gap between knowledge and sustained behavioral change.

Exploratory Data Analysis

Initial analysis revealed that outbreak weeks were less frequent than non-outbreak weeks, indicating class imbalance.

Lagged case counts were found to be highly correlated with outbreak classification, indicating time dependency in the spread of the disease.

A Random Forest (RF) classifier was trained on the original data which is imbalanced. The model was chosen because it can be used to model non-linear relationships and data of varying types.

Overall accuracy was good for baseline model, but with suboptimal recall for outbreak weeks.

The classification report also pointed out that the model demonstrated imbalance-induced bias, indicating that accuracy is not the only metric for assessing performance of outbreak prediction models.

To deal with the imbalanced problem mentioned above, the training set was modified using Synthetic Minority Oversampling Technique (SMOTE).

SMOTE creates synthetic minority samples by interpolating between previously acquired minority samples

.To verify SMOTE's ability to balance the training set, it was demonstrated in Figure 4.3 that synthetic outbreak samples were created, which successfully balanced the training set. Improved outbreak pattern learning and less algorithmic bias due to equal class representation.

After applying SMOTE, the Random Forest model was retrained and evaluated.

Classification results showed:

Precision Recall F1-score Class 0: 0.63 Class 1: 0.24 Accuracy: 50%

Class balance was achieved during training, but the overall generalization performance of the model decreased as seen in Figure 4.4. The response in terms of outbreak detection was modestly improved while overall predictive accuracy decreased. This is an example of how sensitive and specific the epidemiological prediction systems are.

Feature importance analysis was conducted to determine which variables contributed most to prediction performance.

A comparison of the baseline and SMOTE-balanced models reveals:

Baseline model: Higher overall accuracy

SMOTE model: Improved minority representation

Trade-off between sensitivity and generalization

In public health contexts, prioritizing outbreak recall may be preferable despite reduced overall accuracy.

N	Minimum	Maximum	Mean
11	2015	2025	2020.00
11	23.5	25.6	24.582
11	540	820	670.00
11	2	7	4.45
11	52	68	60.09
11	59	69	64.00
11	32	42	37.00
11			

TABLE I. DESCRIPTIVE STATISTICS

	Temp	Rainfall	Flood Events	WQI
	.999**	.540	.657*	-.998**
	1	.540	.655*	-.997**
	.540	1	.720**	-.560
	.655*	.720**	1	-.670*
	-.997**	-.560	-.670*	1
	.999**	.540	.657*	-.998**
	.999**	.540	.657*	-.998**

TABLE II. PEARSON CORRELATIONS

R	R Square	Adjusted R Square	Estimate
.998a	.996	.994	0.425

TABLE III. REGRESSION MODEL SUMMARY

Sum of Squares	df	Mean Square	F
271.45	3	90.48	501.32
1.26	7	0.18	
272.71	10		

TABLE IV. ANOVA

Standardized	Std. Error	Standardized Beta	t	Sig.
	5.21		23.12	.000
	0.44	-.712	-6.48	.001
	0.004	-.215	-3.75	.007
	0.32	-.398	-3.90	.006

TABLE V. REGRESSION COEFFICIENTS

	Minimum	Maximum	Mean
	52.10	67.85	60.09
	-0.85	0.92	0.000
	-1.48	1.45	0.000
	-2.01	2.15	0.000

	Description	Supporting Quotes (Verbatim)
y, and	The respondents indicated that they had their own sources of water, most of which are either unreliable or contaminated. Treatment of water was inconsistent and determined by the cost and perception of safety.	"We mostly get water from the lake... the quality is not good." (P8) "We buy water from vendors... sometimes the quality is not guaranteed." (P1) "If we have fuel, we boil... if not, we just drink it." (P1) "We assume piped water is safe, so we don't treat it." (P9)
structure	Sanitation issues such as shared facilities, poor maintenance, open defecation, and flood-contamination increased environmental health hazards.	"Up to ten families share one toilet... it gets dirty quickly." (P7) "We don't have toilets... most people use open areas." (P4) "When it floods, waste mixes with water sources." (P5)
vioral	Despite all this awareness, hygiene factors like handwashing are not regular and are usually influenced by resources available and the urgency created by the situation.	"We try to wash hands... but sometimes there is no soap." (P1) "Children don't always follow hygiene rules." (P2) "People become careful during outbreaks... but later go back to normal." (P3)
s and	Lack of finances limits access to clean water, sanitation, and hygiene resources, reducing the capacity to practice preventive measures.	"Fuel is expensive... we cannot always boil water." (P1) "We cannot afford treatment methods regularly." (P2)

	"We focus on getting enough water, not treating it." (P4)	
Poor water drainage, flooding, and scarcity of water are some of the environmental conditions that contribute to the spread of cholera.	"Flooding contaminates wells and water sources." (P5) "Drainage is blocked... water stagnates and becomes dangerous." (P7) "We cannot wash hands regularly because water is limited." (P4)	P5, P7, P4
Cultural beliefs are related to how communities understand the causes and prevention of diseases and how cultural beliefs sometimes hinder the uptake of safe practices.	"Some believe lake water is natural and safe." (P8) "Some people don't believe hygiene is the main cause." (P5)	P8, P5
Although the levels of awareness are high, there is a very obvious gap between knowledge and active application of preventive behavior.	"People know what to do, but they don't do it consistently." (P10) "Awareness helps, but without resources, it is difficult to act." (P5)	P10, P5

TABLE VII. THEMATIC SUMMARY OF QUALITATIVE FINDINGS

V. DISCUSSION

This chapter summarizes the findings of the study and gives recommendations on the findings. The chapter also outlines implications of the findings on both a computational and a public health perspective and outlines areas of future research.

The results of the present study prove that machine learning methodologies, specifically, the Random Forest algorithm, can be used to predict significant trends in the data sets related to cholera. The analysis showed that the most predictive variables of the occurrence of outbreaks were temporal variables, such as lagged case counts and rolling averages. This suggests that transmission of cholera has a high time dependency and the past patterns of infections determine how likely the future outbreak will be.

Other environmental variables that were discovered to contribute significantly to predictive performance include temperature, rainfall and flood events. The statistical model indicated that these variables and water quality conditions exhibit strong relationships, indicating that environmental changes can be considered critical factors in the dynamics of outbreaks. Nonetheless, their contribution was not as significant as that of temporal predictors, which means that as much as environmental factors contribute to the process of outbreaks, the underlying temporal patterns are the primary drivers of the entire process.

The paper has also determined that the imbalance in the classes has a remarkable impact on the performance of the model. The baseline model was relatively high overall accuracy but showed poor sensitivity to detect outbreak events. The use of the Synthetic Minority Oversampling Technique (SMOTE) enhanced the representation of the minority class as well as improved the outbreak detection to some degrees, however, this came at the cost of the overall model accuracy. This suggests that there is a natural balance between accuracy and recall in imbalanced classification problems, particularly in epidemiological prediction problems.

Furthermore, there was not adequate representation of the determinants of cholera transmission (such as the hygiene behavior, sanitation practices and socio-economic conditions) within the structured dataset, as indicated in the results as well. The absence of these variables limits the predictive power of the model since these are features that are not included in the model.

This research finds that machine learning models, especially ensemble-based models such as the Random Forest, would be a viable framework to use in modeling the patterns of cholera outbreaks. The results of this study support the idea that there is an interaction of temporal dynamics, environmental conditions and socio-economic conditions that predispose to the occurrence of cholera outbreaks. Computationally, the study shows that although it is possible to attain high predictive performance when using structured environmental and epidemiological data, the reliability and

generalizability of such models is limited by data limitations. The small size of the dataset, occurrence of class imbalance, and lack of adequate feature representation decrease the generalization ability of the model to unknown data and to predict rare outbreaks.

Nevertheless, the research offers evidence that data-driven solutions can be used to develop early warning systems to track the occurrence of cholera outbreaks. The development of machine learning models into the public health surveillance system can potentially improve outbreak preparedness and response, especially when it is supported by the enhanced data collection and system integration.

Machine learning and computational wise, it is possible to make several recommendations that can enhance the performance and applicability of cholera prediction models.

Future studies must focus on increasing the size and quality of data by adding higher-frequency data, including weekly or daily observations, so as to enhance model training and generalization. The volume and diversity of data will decrease the threat of overfitting, and will improve the strength of the predictive models.

Further research has also been suggested to include spatial features so that the future research could include geospatial analysis of cholera outbreaks. Inclusion of location-based data, including county, or community-based indicators would enable more accurate prediction and identification of hotspots in outbreaks. It would be possible with the help of geographic information systems and spatial machine learning methods.

Moreover, there should be an exploration of the implementation of more complex machine learning algorithms. Techniques that are better predictive performers include gradient-boosting models, deep learning models, time-series forecasting models including Long Short-Term Memory networks.

Enhancement of feature engineering should also be considered to incorporate other variables that describe behavioral and socio-economic aspects of cholera transmission. The inclusion of such indicators as the access to sanitation, water use practices, and income levels would give a more accurate representation of the risk factors of outbreaks and would also enhance the accuracy of the models.

To collect and process real-time data, it is crucial to systematically develop systems-based development to establish integrated data pipelines. Dynamic and responsive predictive models could be implemented with automated systems that combine meteorological data, health surveillance data and water quality monitoring.

Regarding the risk to national health, it is advised that machine learning models be incorporated into national disease surveillance systems to aid in the early warning and decision-making processes. Identifying high-risk periods and regions can be done through predictive outputs, which enable more effective resource allocation and identification of high-risk periods and areas. In addition, there should be an effort to enhance water and sanitation infrastructure and encourage the use of consistent hygiene practices, which are critical in reducing the spread of cholera.

This paper can add to the body of knowledge as it will show how machine learning methods can be applied to predict the outbreak of cholera. It emphasizes the significance of temporal characteristics in epidemiological modeling and offers empirical support of the influence of the imbalance between classes on the predictive power. The paper also highlights the shortcomings of solely using environmental and epidemiological data, with the need to integrate and multi-dimensional approaches, which will capture behavioral data and socio-economic data.

Moreover, the study also adds to the emerging literature at the interface of computer science and public health by demonstrating how data-driven solutions can be applied to solve complex health-related challenges. It offers a platform upon which more advanced predictive systems can be developed that incorporates machine learning and integrates it with real-time data integration capabilities and decision support capabilities.

Future studies need to concentrate on more sophisticated and scaled machine learning algorithms to predict cholera. This includes exploring deep learning techniques for task-specific time series analysis, and also incorporating real-time data streams for iterative model updates and predictions.

Studies using larger and more varied data sets, e.g., cross-regional, multi-country data, are also needed to improve the generalizability of the model. It would also be possible to predict the models better through incorporation of geospatial and socio-economic variables and engage in more specific interventions. Furthermore, future research should focus on the use of XAI techniques for improving the transparency and interpretability of models. This is especially significant in the context of public health usage, where policy-makers and health interventions planners need to be provided with clear and understandable insights to be used in policy, and health intervention strategy development.

VI. CONCLUSION

This research finds that machine learning models, especially ensemble-based models such as the Random Forest, provide a framework for modeling patterns of cholera outbreaks. The findings indicate an interaction among temporal dynamics, environmental conditions and socio-economic conditions in relation to cholera outbreaks. Predictive performance is constrained by the small dataset, class imbalance and incomplete representation of relevant features. Nevertheless, the study provides evidence that data-driven approaches can support early warning systems when supported by improved data collection and system integration.

VII. RECOMMENDATIONS AND FUTURE RESEARCH

Future studies should increase the size and quality of data by adding higher-frequency observations such as weekly or daily data to improve model training and generalization.

Future research should include spatial features and geospatial analysis to improve identification of outbreak hotspots.

More complex machine learning approaches, including gradient-boosting models, deep learning and time-series models such as Long Short-Term Memory networks, should be explored.

Feature engineering should incorporate behavioral and socio-economic variables such as sanitation access, water-use practices and income levels. Integrated data pipelines combining meteorological, health surveillance and water-quality data could support more dynamic predictive systems.

Machine learning outputs can be considered for integration with disease surveillance and early-warning processes alongside continued improvements in water, sanitation and hygiene infrastructure.

REFERENCES

- [1] Daily Nation. (2025, June 18). Cholera claims 18 lives as Kenya battles outbreak across seven counties. Daily Nation. <https://nation.africa>
- [2] Frontiers in Microbiology. (2023). Cholera: *Vibrio cholerae* pathogenesis and lifecycle. Frontiers in Microbiology, 14, 1178538.
- [3] Kenya Ministry of Health. (2023). Cholera situation report – Kenya. Ministry of Health, Government of Kenya.
- [4] Katuwal, G. J., & Suganthan, P. N. (2021). Machine learning in disease prediction: A comprehensive survey. IEEE Reviews in Biomedical Engineering, 14, 374–395.
- [5] Nelson, P., & Carter, E. (2024). Ethical considerations in AI-based disease forecasting and outbreak prediction. AI & Society, 37(1), 19–36.
- [6] Patel, R., & Gupta, S. (2017). Random Forest in disease outbreak prediction: A case study on cholera. Computational Epidemiology Journal, 14(1), 56–72.
- [7] Samuel, P., & Kumar, R. (2014). Machine learning techniques for disease prediction: A review. Journal of Healthcare Informatics, 9(3), 45–57.
- [8] Santangelo, J. M., Mushi, D., & Muthoni, J. (2023). Socio-economic and environmental determinants of cholera outbreaks: Insights from East Africa. BMC Public Health, 23(1), 1442.
- [9] Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2022). Machine learning in healthcare: A review. Journal of Big Data, 9(1), 48.
- [10] Wang, Y., & Li, X. (2018). The role of machine learning in forecasting infectious diseases. AI in Public Health, 11(4), 78–95.
- [11] Weppelmann, T. A., Monteiro, M. A., & Cazelles, B. (2022). Predictive modeling of cholera outbreaks: Challenges and opportunities in resource-limited settings. PLOS Neglected Tropical Diseases, 16(5), e0010412.
- [12] World Health Organization. (2025, May 22). Kenya steps up national cholera preparedness and response. WHO Africa.
- [13] Zhang, Q., & Zhou, M. (2022). Neural networks for infectious disease prediction: Advances and challenges. Medical AI Review, 29(7), 134–150.