Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

DATA-DRIVEN MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF CHOLERA OUTBREAK IN WEST AFRICA

***[1]Onyijen, O. H., [2]Olaitan, E.O, [3]Olayinka, T. C., [1]Oyelola, S.**
[1]Department of Mathematical and Physical Sciences, Samuel Adegboyega University, Ogwa, Edo State, Nigeria.
[2]Department of Computer Science and Engineering, University of Hull, United Kingdom
Email: greatolaitaneben.oe@gmail.com
[3]Department of Information Technology and Cybersecurity, Wellspring University, Benin City, Edo State, Nigeria.
***Corresponding Author: ojei.onyijen@gmail.com**

ABSTRACT
The cholera epidemic remains a public threat throughout history, affecting vulnerable populations living with unreliable water and substandard sanitary conditions. Various studies have observed that the occurrence of cholera has a strong linkage with environmental factors such as climate change and geographical location. Poor Hygiene has been strongly linked to the seasonal occurrence and widespread of cholera through the creation of weather patterns that favour the disease's transmission, infection, and the growth of Vibrio cholerae, which cause the disease. Over the past decades, there have been great achievements in developing epidemic models for the proper prediction of cholera. However, machine learning techniques have not been explicitly deployed in modelling cholera epidemics due to the challenges that come with its datasets, such as imbalanced data and missing information. This paper explores the use of machine learning algorithms such as decision tree, random forest, and logistics regression to evaluate the prevalence of cholera epidemics in West African countries while overcoming the data imbalance problem. In addition, mean square error, mean absolute error, F1 score, precision and balanced accuracy metrics were used to evaluate the performance of the three (3) models. The results show that logistic regression has an accuracy of 0.47%, random forest 0.978% and the most efficient model was the decision tree 0.998% with a mean squared error and mean absolute error of 0.001% respectively shows that the model will accurately predict cholera outbreak in Africa. Overall results will improve the understanding of the significant roles of machine learning techniques in healthcare data. The study recommends a review of healthcare systems to facilitate quality data collection and deployment of machine learning techniques.

Keywords: *cholera, machine learning, dataset, algorithm, random forest.*

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

1. Introduction

Ingesting food or water tainted with toxic strains of Vibrio cholera serogroup O1 or O139 results in the acute watery diarrhea sickness known as cholera (Sharmila et al., 2016). If rehydration therapy is not given right away, cholera infection can also be characterized by a rapid loss of fluids, which could lead to severe clinical sequels like lethargy, unconsciousness, confusion, a drop-in blood pressure, and death (Sharmila et al., 2016). Cholera infection is frequently characterized by a rapid onset of watery diarrhea, with or without vomiting, and extensive dehydration (Microbiology Society, 2016). Although untreated cholera can have case fatality rates (CFRs) as high as 30–50%, rehydration therapy has been demonstrated to be beneficial in lowering CFRs to as little as 1%. (Ali et al., 2015). According to the World Health Organization (2019), cholera might cause up to 95,000 deaths worldwide each year (21,000–143,000), disproportionately impacting the most vulnerable and underprivileged people who lack access to effective prevention measures (Babatimehin et al., 2017).

Currently, the seventh pandemic, which originated in Indonesia in 1961 and spread to Africa in 1970, is affecting the entire world (Griffith et al., 2006). It has since been observed that other subregions of Africa have experienced cholera epidemics recurrently (Ajayi & Smith, 2019). Over the last four decades, 3,221,207 cases have been documented and 202,456 deaths have been recorded (Mengel et al., 2014). In the African region, the cholera-prone Member States have established prevention and control strategies and are coordinating the execution of crucial measures to control cholera. However, sub-Saharan Africa has the greatest burden and impact of the disease. Cholera outbreaks frequently affect countries in West Africa. There are places where the disease is spreading endemic-style and others where it is spreading epidemic-style. Several predisposing factors contribute to the disease's occurrence. Low socioeconomic status groups are primarily affected by cholera. The most at-risk groups are those who live in crowded regions with insufficient access to safe water and sanitary facilities, particularly uncontrolled mass urbanization. Every year, cholera affects about 2.9 million people and kills 95,000 people worldwide, mostly in low- and middle-income nations. Forty million (40) million people reside in cholera-endemic areas in Africa alone, where outbreaks are a constant risk (Ali et al., 2012; Hassan, 2021).

In 1991, there were no further cases reported. Two waves of cholera outbreaks hit Kenya in 2022, the first of which hit three counties—Nairobi, Kisumu, and Kiambu—from May through June. Fourteen (14) countries have been impacted by the wave, which was confirmed on October 8 2023. (CDC, 2023). A cholera outbreak occurred in Chegutu, Zimbabwe's urban and peri-urban districts in 2018. Zanzibar experienced a cholera outbreak in 2015–16, with over 4000 cases being documented. Zambia announced a cholera outbreak on October 6, 2017. In the year 2015, there was cholera outbreaks in Tanzania, especially in Dar es SalaamIdai and Kenneth, two powerful cyclones that struck Mozambique in March and April of this year,

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

caused cholera epidemics that resulted in nearly 7,000 cases and 8 fatalities. In Nigeria, there have been cholera outbreaks with high CFRs: in 2010, there were 41,787 cases and 1,716 deaths (or 4.1%). (Worldwide Task Force to Control Cholera, 2010). To reduce cholera-related mortality by 90% by 2030 in all endemic countries, including Nigeria, the Global Task Force on Cholera Control (GTFCC) and its partners supported and coordinated the implementation of a multi-sectoral approach in all endemic countries in 2017 (Martin et al., 2014). Despite WHO's best attempts to contain the cholera outbreak, since December 2020, additional outbreaks have been recorded in the subregion. Consequently, computer methods must be used to forecast the cholera outbreak in Africa.

Computational techniques like machine learning work best in data analysis, particularly in medical diagnosis for minor, specialized diagnostic issues. Through machine learning, which is an application of artificial intelligence, computer-based systems can automatically learn from their experiences and get better over time without having to be explicitly designed (Demsar et al., 2013). Most machine learning techniques fall under the supervised and unsupervised algorithm categories. Supervised algorithms are used when the data used to train is classed and labeled while unsupervised algorithms are utilized with unlabeled data (Sathya & Abraham, 2013). Building models that can take input data and apply statistical analysis to predict an output while updating outputs as new data becomes available is the fundamental idea behind machine learning (Olivera et al., 2017). Due to its ability to handle data in an inventive manner to accomplish its intended goals, machine learning is now used in a wide range of applications, including quick decision-making, virtual personal assistance, social media services, video surveillance, disease identification and diagnosis, drug discovery, and clinical research (Domingos, 2012; Abuassba et al., 2017).

Machine learning's main goal is to recognize complicated patterns automatically and, depending on the information given, make wise decisions. Algorithms are developed to take some inputs and utilize mathematics and logic to produce a reliable output in the innovation of making machines learn, which is termed machine learning, and its intelligence (Edureka, 2019). All the models in machine learning learn from the past and give predictions based on some dataset. Computer learning methods are widely used in predicting cholera, and they get preferable results. With new advancements in machine learning, the detection of cholera will become very easy and cheap. Many datasets related to cholera are available. Hence, machine learning is a necessity for application in medical diagnosis.

Machine learning technology is still working through issues from decades past when attempting to anticipate the disease from a patient's symptoms and the patient's history. Machine learning technology can be effectively used to solve healthcare challenges. With the aid of machine learning models, we may create models that quickly clean and process data and produce results. By utilizing this approach, doctors will make wise choices regarding patient diagnoses, and in turn, the patient will receive wise treatment, improving the quality of patient healthcare services.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Several cholera predictions models have been developed to forecast outbreaks of the disease. This paragraph describes several such systems and highlights their strengths and weaknesses. One such model proposed by (Young, 2017) utilizes newly collected census data from Haiti to identify potential indicators of cholera, such as specific behaviors and living situations.

Another model, proposed by (Pasetto et al., 2017), uses real-time projections and integrates new information as it becomes available. Additionally, (Yue et al.,2014) developed a model that focuses on the impact of climate factors in the estuary of Pearl River. Their research involved daily collection of climate data at meteorological stations in Guangzhou and Shenzhen, which were then converted to monthly data. The model's parameters, including the water temperature coefficient and coefficient of cholera shifting in the regions, were determined through linear regression.
The present study investigates a novel use of machine learning techniques in predicting cholera outbreaks in West Africa. In particular, it explores the use of random forest classifiers, decision trees, and logistic regression and evaluates their application potential for predicting cholera disease.

2.1 Cholera

Every year, Vibrio cholerae produces 3 to 5 million cases of cholera, killing 100,000–120,000 people (Ali, 2012). Infection is spread through the consumption of tainted food or water, particularly affecting areas with inadequate sanitation and access to clean drinking water (Kaper, 2015; Charles, 2015). Watery diarrhea and rapid dehydration are indications of the illness, which if left untreated can cause hypotonic shock and mortality within 12 hours of the onset of symptoms (Kaper, 2015; Charles, 2015). Over the past 200 years, the disease has spread widely, including many recent epidemics in Haiti, Vietnam, and Zimbabwe (Chin, 2011; Mason, 2009). In many regions of the world where cholera is present, annual, and seasonal outbreaks also happen. Due to the toxigenic V. cholerae's capacity to thrive year-round in the aquatic environment, endemic regions of the world, including those in Asia, Africa, and the Americas (Alam, 2006; Faruque, 2008). Seasonal outbreaks vary in timing and severity based on several environmental conditions, such as rainfall, salinity, temperature, and plankton blooms (Huq, 2005). Transmission between members of the same family occurs frequently, and epidemics of cholera are frequently made worse in densely populated places with inadequate infrastructure (Weil, 2009).

2.2 Machine Learning Approaches

Using computational techniques to transform empirical data into useable models is the subject of the branch of study known as "machine learning" (Thomas, 2017). Conventional statistics and artificial intelligence communities were the ancestors of the machine learning discipline. Machine learning has emerged during the past ten years as one of the trendiest areas of

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

computational science thanks to the efforts of large firms like Google, Microsoft, Facebook, Amazon, and others. Huge volumes of data have already been acquired through their business processes and will continue to be. This has given rise to a chance to revive statistical and computational methods for automatically creating effective models from data.
Machine Learning Overview 9
Machine Learning Overview
Taiwo Oladipupo Ayodele
X

Machine Learning Overview

Computational statistics, which focuses on making predictions with computers, is closely related to a subset of machine learning, but not all machine learning is statistical learning. The field of machine learning benefits from the tools, theory, and application domains that come from the study of mathematical optimization. Data mining is a related area of study that focuses on unsupervised learning for exploratory data analysis (Bishop, 2006). Machine learning is sometimes known as "predictive analytics" when it is used to solve business challenges. Many medical-related projects now make advantage of the developing science of machine learning. All machine learning models make predictions based on some dataset and take historical data into account. The identification of cholera will be made exceedingly simple and affordable by recent advances in machine learning. There are a lot of cholera-related datasets out there. Considering this, machine learning is required for applications in medical diagnosis. Predicting the likelihood that a patient may develop cholera is our goal. Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning are the four primary divisions of machine learning techniques (Mohammed, 2016).

Supervised Learning can be classified into two (2) categories of algorithms: Classification and Regression algorithms. The classification algorithms are used to make predictions. They include Neural Networks, K-NN, Decision Trees, Random Forests, Support Vector Machines, Naïve Bayes etc. Regression is a technique that is used to predict continuous quantity output (Gitau, 2018). Regression is also a method for predicting, forecasting, and determining correlations between quantitative data (Mark *et al.*, 2015). Unsupervised learning, on the other hand, is the process of training a system or machine with unclassified or unlabelled data and allowing the system to generate predictions without being supervised. In this case, the system groups unsorted data based on similarities, patterns, and differences on its own, without the need for any training (Shubham, 2020). Clustering and Association are the two categories of unsupervised learning. Unsupervised learning is classified into two (2) categories: Clustering and Association.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

2.3 Review of related literature

In the study by Giardina et al. (2006) titled "Supervised learning approach to predicting coronary heart disease complications in type 2 diabetes mellitus patients," a supervised machine learning (ML) approach was developed to predict and classify DM type 2 according to the presence or absence of coronary artery disease complications. The supervised ML predictive model for acute ischemic stroke post intra-arterial therapy was built in the article of Asadi et al. (2014) on "Machine learning for outcome prediction of acute ischemic stroke following intra-arterial therapy." The study further developed a robust learning model that may be able to optimize the choices of medical treatment and endovascular activities in the management of acute strokes. The model demonstrated a promising accuracy of prediction. In the study of Dai et al. (2015), supervised ML algorithms were used to create a predictive model for hospitalization due to heart disease. The predictive supervised ML model for the prediction of post-induction hypotension was constructed based on the paper of Samir et al. (2018) on "Supervised Machine Learning Predictive Analytics for Prediction of Postinduction Hypertension." The study's findings indicated that the ability of supervised ML models for predictive analytics in the field of anaesthesiology is demonstrated by the success observed in post-induction hypotension prediction. On a dataset of cholera outbreaks for Indian coastal districts from 2010 to 2018, a Random Forest classifier model is developed, trained, and tested. The random forest classifier model accurately recognizes 89.5 percent of outbreaks with an Accuracy of 0.99, an F1 Score of 0.942, and a Sensitivity score of 0.895.

The model's outputs showed spatial temporal trends that were correlated with seasons and coastal locations. The dataset used for the model development in Rajagopalan & Vollmer's (2019) study on "Rapid detection of heart rate fragmentation and cardiac arrhythmias cycle-by-cycle analysis, supervised machine learning model and novel insights" was obtained from an urban hospital in Boston. Five models were created using the likelihood ratio test, SVM, AdaBoost, LR, and Naive Bayes algorithms. A supervised ML model was created for the quick detection of heart arrhythmias and heart rate fragmentation. A prediction model was created using a random forest algorithm and a dataset of 300 cases of arrhythmic, non-arrhythmic, and those without any clinically relevant cardiac problems. 104 independent instances were used to evaluate the model, which was quite effective. Using crucial climate variables derived from atmospheric, terrestrial, and oceanic satellites, the authors (Daisy et al., 2020) present a novel exploration of the potential of a machine learning method to forecast environmental cholera risk in coastal India, which has a population of more than 200 million people. In the study of Rustam et al. (2020) on "COVID-19 future forecasting using supervised machine learning models". The least absolute shrinkage and selection (LASSO), exponential smoothing (ES), linear regression (LR), and support vector machine (SVM) techniques were used to build the model. The study showed that the algorithms for supervised machine learning could estimate how many patients would be afflicted by COVID-19 in the future.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

In the study by Young (2017) put forward a proposal for a research paper aimed at predicting cholera positive cases in Haiti through the use of recently collected census data. The goal of the study was to determine if certain behaviours and living situations could be utilized as indicators for identifying individuals with cholera. The research is an innovative approach to predicting cholera in Haiti, which has the potential to significantly impact public health in the region. The study's focus on determining if certain behaviours and living situations can be used as indicators for identifying individuals with cholera could provide valuable insight into the disease's spread and inform future prevention strategies. In the study by Pasetto et al. (2017) explored real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. Their study tested a real-time forecasting framework that could readily integrate new information as soon as it became available and periodically issue updated forecasts. Similarly, the real-time forecasting framework proposed in this study has the potential to revolutionize the way cholera outbreaks are managed. By integrating new information as soon as it becomes available and issuing updated forecasts, the framework could help authorities take proactive measures to prevent the spread of cholera. In the study by Yue et al. (2014) introduced a cholera prediction model that focused specifically on the effect of climate factors in the estuary of Pearl River. These research work involved the daily collection of climate data at meteorological stations in Guangzhou and Shenzhen, with daily data being converted to monthly data. Parameter values, such as the water temperature coefficient and the coefficient of cholera shifting in the regions, were determined through linear regression. Finally, the cholera prediction model introduced in this study provides a unique perspective on the impact of climate factors on the disease's spread. By collecting daily climate data, the researchers were able to generate monthly data sets and determine parameter values that could help predict cholera outbreaks. The model's focus on the estuary of Pearl River makes it particularly relevant to regions with similar geographic characteristics. Previous methods for predicting cholera outbreaks have been limited in various ways, such as the use of a small number of features for prediction and the need to wait for the reporting of certain cases before obtaining data.

However, recent research has proposed new approaches to cholera prediction that are based on rainfall and seasonal weather changes. For example (Pasetto et al., 2017) have developed a cholera prediction model that takes into account rainfall patterns, while (Leo, 2020) has proposed a model that uses seasonal weather changes as a predictor of cholera outbreaks. Previous cholera prediction models have used fewer features for prediction, leading to lower accuracy. As a result, further research is necessary to address this shortcoming. In this study, a more extensive set of cholera features was employed to predict outbreaks before they occur, utilizing the Random Forest, Decision Tree, and Logistics Regression Classification techniques. Furthermore, implementation was completed using Python programming to address the gaps identified in previous studies.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

3.1 Dataset preparation

The training set for the physical examination data consists of nine hundred and eighty-four (984) instances. The information includes five (5) physical examination indexes: the nation, the year, the number of cholera cases reported, the number of cholera deaths reported, the cholera case fatality rate, and the WHO region. Anaconda software and Jupyter Notebook were used to conduct the analysis. The datasets were initially saved as Microsoft Excel documents. However, they were then pre-processed and prepared in comma-separated values file (CSV) format. The predicted supervised machine learning models in this work were created using the Python programming language.

3.2 Training Dataset Using Python

The Python script was utilized for prediction on a Jupyter Notebook within the Anaconda environment. Anaconda functions as a package manager, environment manager, and distribution of Python, encompassing numerous open-source packages. This proves advantageous for data science projects. Initially, we imported the NumPy, Scikit-learn, SciPy, and Pandas modules, subsequently loading the cholera datasets. Following this, we assessed dataset balance and executed a sampling procedure to counterbalance the dataset. The training data was then utilized to construct models including Random Forest, Decision Tree, and Logistics Regression. We evaluated the performance of these models using the testing data. Lastly, we performed evaluation metrics to identify the optimal performing models or algorithms.

1. Logistics Regression

It is a potent supervised machine learning approach utilized for binary classification issues is logistic regression. The best approach to conceptualize logistic regression is as a linear regression applied to classification issues. In essence, logistic regression models a binary output variable using the logistic function described below (Tolles & Meurer, 2016). Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Logistic regression models a relationship between predictor variables and a categorical response variable. Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors. There are various metrics to evaluate a logistic regression model such as confusion matrix, AUC-ROC curve, etc.

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}} \quad (1)$$

Where, x is the input value, y is the predicted output, b_0 is the bias or intercept term and b_1 is the coefficient for input (x).

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

2. Decision Tree

A well-known non-parametric supervised learning technique is the decision tree (DT). Both the classification and regression tasks are carried out using DT learning techniques (Pedregosa, 2011). The most well-known DT algorithms are ID3, C4.5, and CART. Additionally, Sarker et al recently proposed BehavDT (Sarker, 2019) and IntrudTree (Sarker, 2020) are successful in the pertinent application domains, such as user behavior analytics and cybersecurity analytics, respectively. DT categorizes the cases by arranging the tree from the root to a few leaf nodes, as shown in s. Starting at the root node of the tree and working our way down the branch that corresponds to the attribute value, instances are categorized by inspecting the attribute defined by that node. The most widely used criterion for splitting is "gini" for the Gini impurity and "entropy" for the information gain, both of which have mathematical expressions as (Pedregosa, 2011). Entropy is nothing but the uncertainty in our dataset or measure of disorder. The higher the entropy, the harder it is to draw any conclusions from that information. Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favours larger partitions and easy to implement whereas information gain favours smaller partitions with distinct values. Information gain measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node.

$$\text{Entropy, } E\ (S) = \ -P(+)\ Logp(+) - p(-)\ logp(-)$$

$$E\ (S) = \ \sum_{i=1}^{c} -pi\ log_2 \text{pi} \quad (2)$$

$$\text{Gini index } = 1 - \ \sum_{i=1}^{c} (pi)^2 \quad (3)$$

$$\text{Information gain} = E(S) - E(S|X) \quad (4)$$

3. Random Forest (RF)

A well-known ensemble classification method used in a variety of machine learning and data science applications is the random forest classifier (Breiman, 2001). Random Forest is a technique that uses ensemble learning, that combines many weak classifiers to provide solutions to complex problems. As the name suggests random forest consists of many decision trees. Rather than depending on one tree it takes the prediction from each tree and based on the majority votes of predictions, predicts the final output. The "parallel ensembling" technique used in this method fits multiple decision tree classifiers simultaneously on various data set sub-samples and uses averages or majority voting to determine the conclusion. As a result, it reduces the over-fitting issue and improves prediction and control (Pedregosa, 2011). Consequently, RF learning models with several decision trees often have higher accuracy than models with only one decision tree (Sarker, 2019). Combining bootstrap aggregation (bagging) and random feature selection, it creates a set of decision trees with controlled variance. It fits well for both categorical and continuous variables and can be applied to classification and regression issues.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

$$nl_j = W_j\ C_j - Wleft_{(j)}C_{left(j)} - W_{right(j)}C_{right(j)} \quad (5)$$

The installation of python programming language working environment like jupyter notebook on anaconda was done on the windows 10 operating system. Go to the Anaconda website and choose a Python 3.x graphical installer (A) or a Python 2.x graphical installer (B). if you are not sure which python version you want to install, choose python 3. Note your installation location and then click Next. This is an important part of the installation process. The recommended approach is to not check the box to add Anaconda to your path. This means you will have to use Anaconda Navigator or the Anaconda Command Prompt (located in the Start Menu under "Anaconda") when you wish to use Anaconda. The fundamental hardware requirements for python deployed for the decision tree procedure used in this research should be a workstation or a personal computer (Laptop) with at least the following configuration; Quad-core Processor with a minimum of 2.0GHz, 4GB Random Access Memory (RAM), 500GB Hard Disk Drive (HDD), NVIDIA Graphic Adapter, Relevant input devices such as a mouse, keyboard, and so on. The hardware requirements are solely dependent on the total number of data to be mined. The software requirements used in this project are as follows: Operating System: Windows 10, Python anaconda. We began by importing the essential libraries such as NumPy, Pandas, and the cholera dataset. Pre-processing of the dataset is carried out to identify any abnormalities that may exist within the dataset. Visualization of the dataset is done through the integration of libraries such as Seaborn, Plotly, and Matplotlib. The next step involves dividing the dataset into two parts, the train, and test sets, in an 80:20 ratio. The algorithms used in the process are imported via Scikit-learn. Finally, the performance evaluation of the various algorithms is carried out by importing sciPY. This process allows for the identification of the best algorithm for analysing the cholera dataset.

3.3 Feature Selection

The feature selection was carried out on the dataset. The dataset contains five (5) attributes such as country, year, number of reported cases of cholera, Number of reported deaths from cholera, Cholera case fatality rate and Who region. These attributes were selected for the training and used for the models. Data pre-processing is putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and that the ordering does not affect the learning process. Cleaning the data to remove unwanted data, missing values, rows, columns, duplicate values, and data type conversion. You might even have to restructure the dataset and change the rows and columns or index of rows and columns. Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training, then the machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. After training your model, you must check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy. When used on testing data, you get an accurate measure of how your model will perform and its speed, the model looks for the patterns in the data, and then we ask it to make predictions.

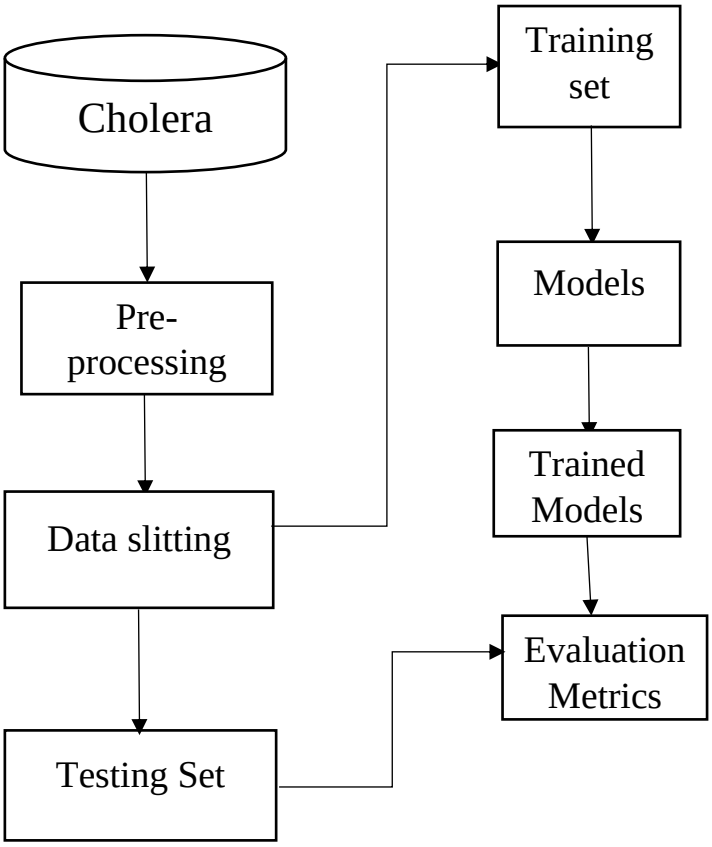


Figure 1: Detailed Architecture for prediction of cholera

4.0 Results and Discussion

This section covers the discussion of the result of the different machine learning model performance.

4.1 Performance Analysis

The study was conducted to analyze cholera cases and death rates in Africa. The data from all cholera outbreaks in each African country from 1970 to 2016 was gathered to predict the best machine learning model for cholera outbreak in Africa. Previous research had demonstrated the utility of machine learning models in predicting cholera outbreaks globally, but not specifically in Africa. The dataset was subjected to three machine learning classifier methods: Logistics Regression, Random Forest, and Decision Tree, to determine the best performance

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

for cholera prediction. The Decision Tree model exhibited the best accuracy of 0.99% and could be utilized to assess the death rate and number of cholera cases.

To affirm the Decision Tree model's accuracy, the mean absolute error (MAE) of 0.001% and mean square error (MSE) of 0.001% were evaluated and shown in Table 2. The Random Forest model showed a good accuracy value of 0.98%, with a mean absolute error (MAE) of 0.124% and a mean square error of 5.952% (MSE) in comparison to Logistics Regression, as depicted in Table 2. The data presented in Figure 2 shows the distribution of cholera cases and death rate in each country. Nigeria had nearly 60,000 cholera cases in 1990. Figure 3 illustrates the cholera death cases in Africa, with the Democratic Republic of Congo having the highest death rate of over 20,000. The correlation plot heatmap in Figure 4 shows a two-dimensional correlation matrix between two discrete dimensions, using colored cells to represent data from a monochromatic scale. The values of the first dimension appear as the rows of the table, while the second dimension is a column. The covariance plot heatmap, as shown in Figure 5, is a measure of how well changes in one variable are associated with changes in a second variable. Specifically, it is a measure of the degree to which two variables are linearly correlated. Table 1 displays the precision, Recall, F1-score, and support. Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

The precision, recall, and F1-score values for the Logistics Regression, Random Forest, and Decision Tree models are presented in Table 1. The results indicate that the Decision Tree model has the highest precision, recall, and F1-score, indicating its superiority over the other models. The support value indicates the number of occurrences of each class in the dataset. In conclusion, the research demonstrates that machine learning models can be effectively utilized to predict cholera outbreaks in West Africa. The Decision Tree model exhibited the highest accuracy, precision, recall, and F1-score, making it the most suitable model for cholera prediction. The study's findings provide valuable insights for public health officials and policymakers, enabling them to take proactive measures to prevent and control cholera outbreaks in Africa. Precision equation is depicted below;

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{6}$$

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive} \tag{7}$$

Recall is the ratio of correctly predicted outcomes to all predictions. It is also known as sensitivity or specificity.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{8}$$

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \quad (9)$$

The F1 score combines these three metrics into one single metric that ranges from 0 to 1 and it considers both Precision and Recall.

$$F1 = 2 * \frac{Precision-Recall}{Precision+Recall} \quad (10)$$

Accuracy is a performance metric for a machine classification model defined as the ratio of true positives to true negatives for all positive and negative observations, In other words, accuracy tells us how often we can expect our machine learning model to correctly predict the outcome of the total number of times it made the prediction.

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+False\ Negative+True\ Negative+False\ Positive} \quad (11)$$

Mean Squared Error (MSE) is calculated by taking the average of the square of the difference between the original and predicted values of the data.

Mean Absolute Error is the sum of absolute errors over the length of observations/predictions.

$$\frac{1}{N}\sum_{i=1}^{n}(actual\ values - predicted\ values)^2 \quad (12)$$

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

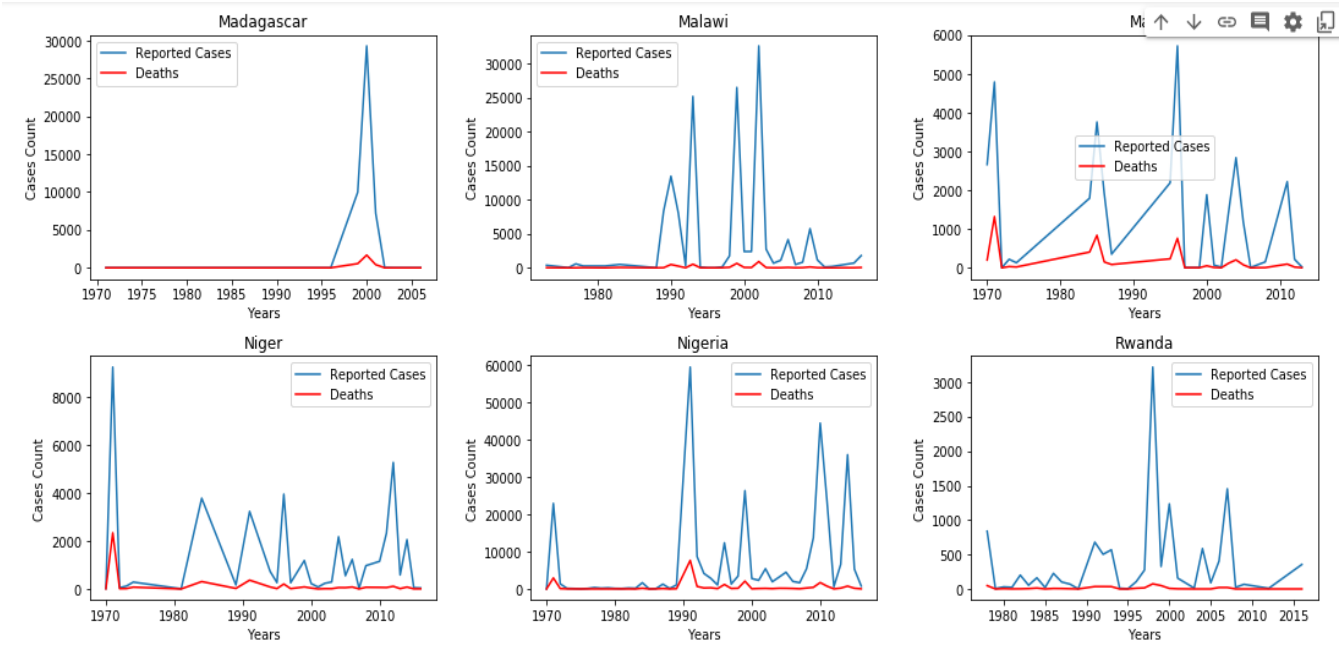
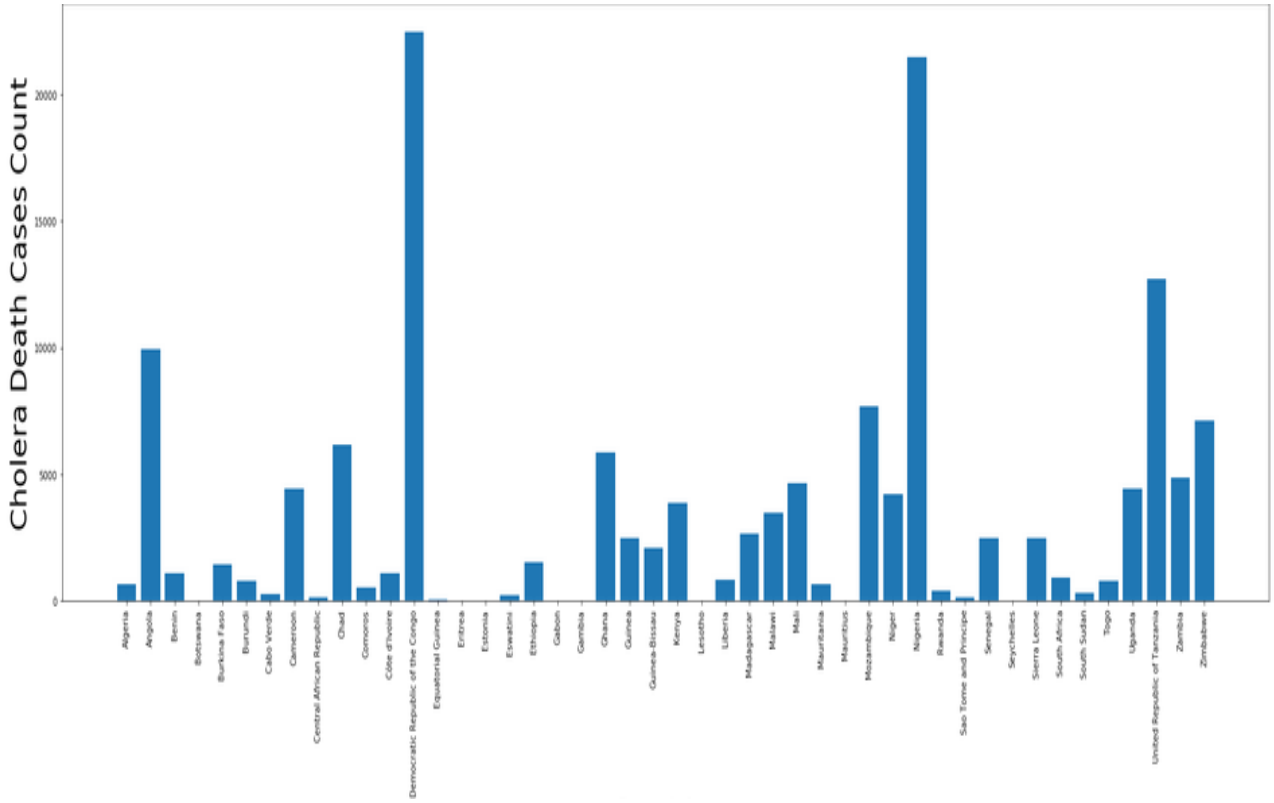


Figure 2: Countries Cholera cases and Death rate distribution



`Figure 3: Distribution of cholera death cases in Africa

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

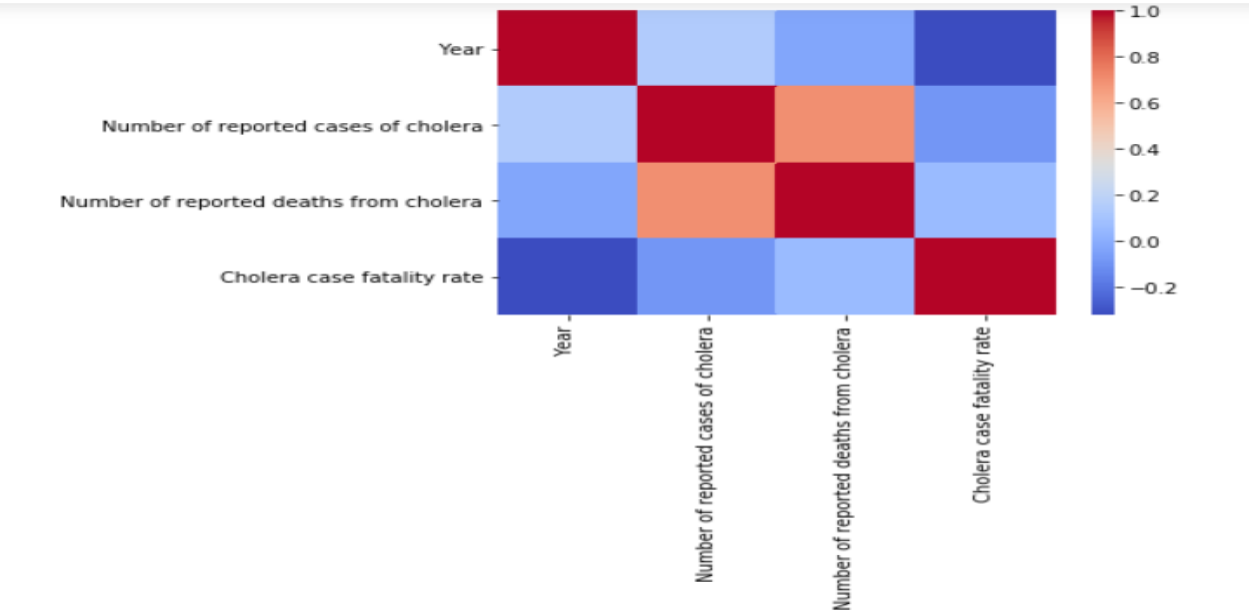


Figure 4: Correlation plot heatmap

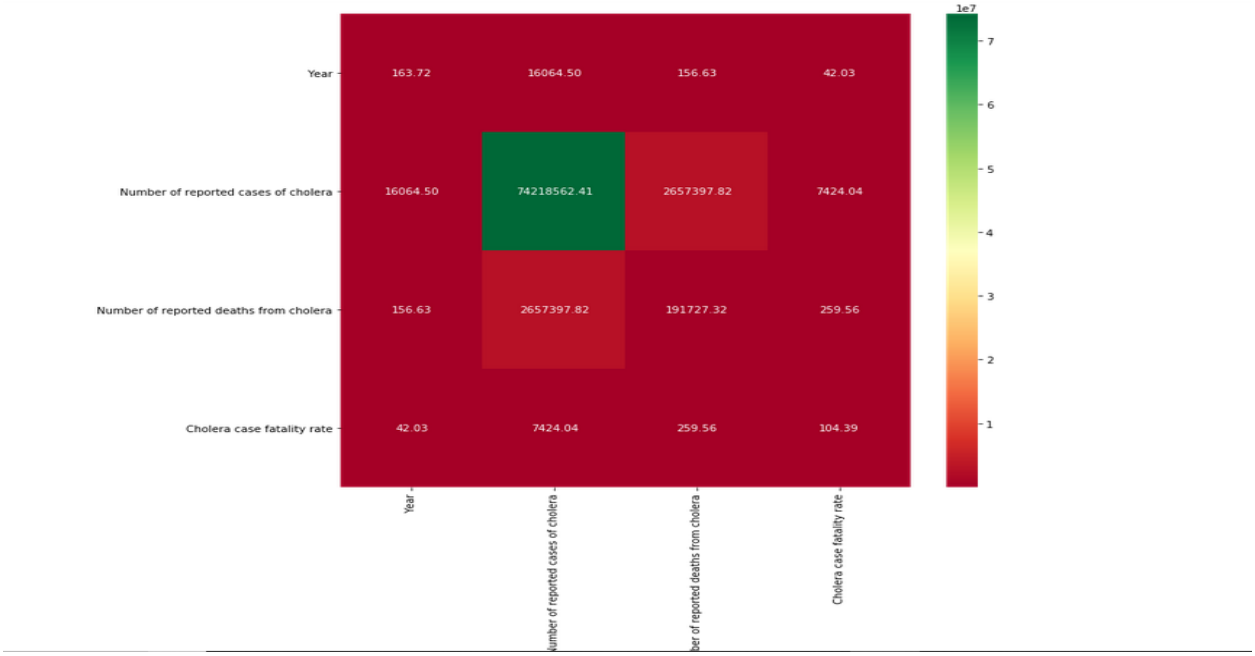


Figure 5: Covariance plot Heatmap

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Table 1: Prediction scores for the machine learning models

Models		Precision	Recall	F1-score	Support
Logistics Regression	0	1.00	0.47	0.64	157
	1	0.45	0.76	0.56	87
Random Forest	0	1.00	1.00	1.00	157
	1	0.99	1.00	0.99	87
Decision Tree	0	1.00	1.00	1.00	157
	1	1.00	1.00	1.00	87

Table 2: Performance evaluation result of the model

Metrics	Logistic Regression	Random Forest	Decision Tree
Accuracy	0.471	0.979	0.998
Mean absolute error	2.096	0.124	0.001
Mean squared error	57.411	5.952	0.001

5.0 Conclusion

The data-driven machine learning approaches have been used for the prediction of cholera outbreak in West Africa with the hope of reducing the menace of this disease by early diagnosis and management. In predicting the cholera outbreak in West Africa three (3) machine learning models such as Logistics Regression, Decision Tree, and Random Forest were used. Decision tree predictive learning-based model was found to be the best model among the developed models with 0.99% in terms of accuracy, mean absolute error 0.001(MSE) and mean squared error 0.001(MSE). This model will help medical personnel predict cholera outbreaks more accurately and it is considered a reliable tool for early detection of cholera outbreak to reduce the effect of the disease.

The model under consideration herein was meticulously designed and comprehensively tested with the aid of a case study dataset. However, it is imperative to note that the accuracy of the proposed model can be further augmented by obtaining additional datasets from a plethora of diverse or multiple case studies. The Decision Tree, Random forest, and Logistics Regression Classification Machine learning algorithms were utilized in proposing the model. It is noteworthy that exploring the possibility of using an alternative ML algorithm is highly recommended to ascertain the accuracy rate, and more benchmarking is warranted. It is highly recommended to employ mixed-method research to gain a profound understanding of the case study and also to mitigate the risk of being biased.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

References

1. Abuassba, A. O. M. Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving classification performance through an advanced ensemble based heterogeneous extreme learning machines. Computational Intelligence and Neuroscience, 1-11.DOI: 10.1155/2017/3405463.
2. Ajayi, A., & Smith, S. I. (2019). Recurrent cholera epidemics in Africa: which way forward? A literature review. *Infection*, 47, 341-349.
3. Alam M, Sultana M, Nair, G. B., Sack, R. B., Sack, D. A., Siddique, A. K., Ali, A., Huq, A., Colwell, R. R. (2006). Toxigenic *Vibrio cholerae* in the aquatic environment of Mathbaria, Bangladesh. *Applied Environmental Microbiology*, 72, 2849–2855
4. Ali, M., Lopez, A. L., You, Y. A., Kim, Y. E., Sah, B., Maskery, B., Clemens, J. (2012). The global burden of cholera. *Bull World Health Organisation*, 90, 209–18. https://doi.org/10.2471/BLT.11.093427.
5. Ali, M., Nelson, A. R., Lopez, A. L., & Sack, D. A. (2015). Updated global burden of cholera in endemic countries. *Journal of PLoS neglected tropical diseases*, *9*(6).
6. Asadi, H., Dowling, R., Yan, B., Mitchell, P., et al. (2014). Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS ONE*. 9(2).
7. Babatimehin, O., Uyeh, J., & Onukogu, A. (2017). Analysis of the re-emergence and occurrence of cholera in Lagos State, Nigeria: *Bulletin of Geography*, 21–32.
8. Breiman, L. (2001). Random forest. *Machine Learn*ing, 45, 5–32.
9. Charles, R. C., & Ryan, E. T. (2011). Cholera in the 21st century. *Current Opinion on Infectectious Diseases*, 24, 472–477
10. Chin, C. S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R, Bullard, J., et al. (2011). The origin of the Haitian cholera outbreak strain. *Nigerian English Journal of Medicine*, 364, 33–42.
11. Dai, W., Brisimia, T. S., Adams W. G., Mela, T., Saligrama, V., & Ioannis, C. P. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. *International Journal of Medical Information*, 84–3, 189–97.
12. Demšar, J., Curk, T., Erjavec, A. et al. (2013). Orange: data mining toolbox in python. *Journal of Machine Learning Research*, 14.
13. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55 (10), 78–87.
14. Giardina, M., Azuaje, F., McCullagh, P., et al. (2006). Supervised learning approach to predicting coronary heart disease complications in type 2 diabetes mellitus patients. *In: 6th IEEE symposium on bioinformatics and bioengineering (BIBE'06), Arlington*. 325–33.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

15. Griffith, D. C., Kelly-Hope, L. A., Miller, M. A. (2006). Review of reported cholera outbreaks worldwide, 1995-2005. *American Journal for Tropical Medicine Hygiene*, 75, 973-977.
16. Harris, J. B., LaRocque, R. C., Chowdhury, F., Khan, A. I., Logvinenko, T., Faruque, A. S. G., Ryan, E, T., Qadri, F., Calderwood, S. B. (2008). Susceptibility to *Vibrio cholerae* infection in a cohort of household contacts of patients with cholera in Bangladesh. *PLoS Negl. Tropical Disease*, 211-221.
17. Hassan, O. B. (2021). COVID-19, infection control, and cholera. Retrieved on 23 January 2023 from https:// www. healt heuropa. eu/ covid- 19- infec tion- contr ol- and- chole ra/ 107561/.
18. Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., & Morris, J. G. (2005). Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. *Applied Environment. Microbiology*, 71, 4645–4654.
19. Kaper, J. B., Morris, J. G., & Levine, M. M. (2015). Cholera. *Clinical Microbiology Review*, 8, 48–86.
20. Martin, S., Lopez, A. L., Bellos, A., Deen, J., Ali, M., & Alberti, K. (2014). Post-licensure deployment of oral cholera vaccines: a systematic review. *Bull World Health Organization*, 92(12), 881–893.
21. Mason, P. R. (2009). Zimbabwe experiences the worst epidemic of cholera in Africa. *Journal of Infection and Development Countries*, 3, 148–151.
22. Mengel, M. A., Delrieu, I., Heyerdahl, L., et al. (2014). Cholera outbreaks in Africa. *Current Top Microbiology Immunology*, 379, 117-144.
23. Microbiology Society (2016). Cholera: Death by diarrhoea. Retrieved on February 24, 2022, from https://microbiologysociety.org/resource_library/knowledge-search/schoolzone-cholera-an-epidemic-in-haiti.html
24. Mohammed, M, Khan, M. B, & Mohammed, B. E. (2016). *Machine learning: algorithms and applications.* CRC Press.
25. Nigeria Centre for Disease Control. (2019). National Strategic Plan of Action on Cholera Control, Abuja. Retrieved on January 15, 2022 from (PDF) Analysis of the Re-emergence and Occurrence of Cholera in Lagos State, Nigeria (researchgate.net)
26. Olivera, A. R., Roesler, V., Iochpe, C., et al. (2017). Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. *São Paulo Medical Journal*, 135(3), 234–246.
27. Pedregosa, F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V, et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research.* 12, 2825–30.
28. Rajagopalan, A., & Vollmer, M. (2019). Rapid detection of heart rate fragmentation and cardiac arrhythmias: cycle-by-cycle rr analysis, supervised machine learning model and novel insights. *Artifcial intelligence in medicine. AIME 2019. Springer, Cham.* 11526.

Western European Journal of Modern Experiments and Scientific Methods
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/1
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

29. Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W., & Choi, G. S. (2020). *COVID-19 future forecasting using supervised machine learning models*. IEEE Access.
30. Samir, K., Prathamesh, K., Andrew, D. R., et al. (2018). Supervised machinelearning predictive analytics for prediction of postinduction hypotension. *Anesthesiology*. 129, 675–88
31. Sarker, I. H., Kayes, A. S. M., Badsha, S., Algahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data, 7(1)*, 1-29.
32. Sarker, I. H., Watters, p., & Kayes A. S. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data, 6(1)*,1-28.
33. Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34–38.
34. Sharmila, T, & Thomas, T. A. (2016). Pathogenesis of cholera: recent prospectives in rapid detection and prevention of cholera, 129–144.
35. Weil, A. A., Khan, A. I., Chowdhury, F., Larocque, R. C., Faruque, A. S. G., Ryan, E. T., Calderwood, S. B., et al. (2009). Clinical outcomes in household contacts of patients with cholera in Bangladesh. *Clinical Infection and. Diseases*, 49, 1473–1479.