Journal of Applied Sciences, Information, and Computing
Vol. 2, No. 2 (December 2021)
School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://jasic.kiu.ac.ug

CHOLERA PREDICTION MODEL USING FEATURE CLUSTERING BAYESIAN TECHNIQUE

*Ya'u Nuhu[1], Yusuf Musa Malgwi[2], Adamu Abdullahi Garba[3], Usman Muhammad Bala[4]

[1]Department of Computer Science, Federal Polytechnic, Damaturu, Yobe State, Nigeria, yaunuhu@fedpodam.edu.ng
[2]Department of Computer Science, Moddibo Adama University, Yola, Nigeria, yumalgwi@mautech.edu.ng
[3,4]Department of Computer Science, Yobe State University, Damaturu, Nigeria, adamugaidam@gmail.com, usman@ysu.edu.ng

Abstract

Cholera is one the most deadly disease that is mostly caused due by poor sanitation or and drinking contaminated water or food with a bacterium called Vibrio Cholera. Many researchers have provided a solution to prevent the outbreak of cholera using various methods, the recent ones are using machine learning techniques and some mathematical methods such as mathematical epidemiological, spatial statistics, and based on association rule mining using the nonstandard distribution dataset to mention a few. These few methods are mostly used in predicting cholera outbreaks but have some limitations, such as using fewer features for prediction, waiting until certain cases are reported before getting data, based on Rainfall, based on the flowing speed of rivers, wind direction, and flood, etc. in this research a more comprehensive cholera features would be used in predicting an outbreak before it occurs based on the existing secondary dataset using The Naïve Bayesian Classification technique. The proposed model has more features and is not dependent on certain events to occur before predicting any outbreak. Python programming was used in implementing the proposed model. An accuracy of 99% was achieved and it shows it is better than the previous models used in predicting cholera outbreaks.

Keywords: Cholera, Bacterium, Clustering, Algorithm, Prediction, Bayesian, Python

1. INTRODUCTION

Cholera remains a major risk despite global institutional attention. Cholera is a rapidly-dehydrating diarrheal disease, caused by toxigenic serogroups of the bacterium Vibrio cholera; the disease is closely associated with poverty, poor sanitation, and lack of clean drinking water. Historically, devastating outbreaks of cholera resulted in millions of cases and hundreds of thousands of deaths. Currently, cholera remains an important public health problem in many countries, occurring as an endemic disease in some regions and causing major

Epidemics in some low and middle-income countries (Santé, & World Health Organization, 2017).

A cholera outbreak is defined by the occurrence of at least one confirmed case of cholera and evidence of local transmission. Outbreaks can also occur in areas with sustained (year-round) transmission and are defined as an unexpected increase (in magnitude or timing) of suspected cases over 2 consecutive weeks, of which some are laboratory confirmed. Adequate disease surveillance is critical in ensuring early detection of outbreaks (Santé et al., 2017)

Recently, Cholera has attracted the attention of the media in Nigeria. The estimated global burden of the disease is about 4.3 million cases and 143,000 deaths per year of which Africa shares a greater chunk. The recurrent outbreak nature of cholera disease in

endemic countries most especially Nigeria makes the disease a major health problem (Salako, Ajayi & Smith, 2021).

In addition, cholera is one of the primary causes of morbidity and mortality, with incidence occurring in both small outbreaks and large epidemics. The transmission of cholera in Nigeria might be facilitated by numerous factors such as lack of access to safe drinking water, unhygienic environment, environmental disasters, literacy level, population congestion, and internal conflicts which may result in population displacement to Internally Displaced Persons (IDP) camps. The provision of safe drinking water remains a serious issue of concern and this necessitates people even in cities to buy street vented water which has a high risk of being contaminated (Leckebusch & Abdussalam, 2015).

(Shirzad, Ataei & Saadatfar, 2021) states that the need for a prediction model arises due to the trend of the application of data mining in healthcare. Data mining is the modern way of discovering knowledge among databases that leads to statistical analysis, pattern recognition, and information prediction. Today, one of the most important applications of data mining is in the healthcare field which leads to many advances in this area to increase the effectiveness of treatments, reduce the risks, decrease the costs, better patient relationships, early disease diagnosis, etc.

Nowadays, everything generates data and this huge amount of data existing around the world (which leads to the emergence of big data phenomenon) needs to be refined like oil to discover useful knowledge.

Cholera is a dangerous disease that can lead to the loss of many lives. When a cholera outbreak occurs it results to fear, panic and also affects the economic development of the affected area as a result, business partners may wish to withdraw their activities. Hence, there is a need for a reliable technique that would predict such outbreaks, which will be of great importance to society.

However, in the previous predictions methods used has some limitation which includes, using fewer features for prediction, waiting until certain cases are reported before getting data (Reiner, King, Emch, Yunus, Faruque, & Pascual, 2012), According to Pasetto, Finger, Rinaldo & Bertuzzo (2017) proposed a cholera prediction based on Rainfall. (Leo, 2020) proposed cholera prediction based on based-on seasonal weather changes linkages.

However, based on the previous models fewer cholera features for predictions were used and this would provide less accuracy in cholera prediction. Therefore, more research is needed to overcome these limitations. in this research, more comprehensive cholera features were be used in predicting an outbreak before it occurs based on the existing secondary dataset using the Naïve Bayesian Classification technique, and python programming was used for implementation to fill the above-identified gaps found in the previous studies.

2. Research Objectives

The following are the objectives of this study;

i. Review the existing frameworks related to cholera prediction and classification models.
ii. Use Python Programming language to implement Naïve Bayes Classification algorithm for the proposed Model.
iii. Test and evaluate the proposed model based on Naïve Bayes Classification Algorithm.

3. Review of Related Literature

Various prediction models have been developed which are used to predict the cholera outbreak. Examples of such systems are described in the paragraph below. Which shows their strength and weakness in their operation.

Young (2017) proposed a research paper on Predicting cholera Positive cases in Haiti using new census data recently collected from Haiti and attempting to predict if certain behaviors and living situations can be used as an indicator for determining if a person has cholera.

Also, Pasetto, Finger, Rinaldo & Bertuzzo (2017), proposed Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. In this research, a real-time forecasting framework was tested that readily integrates new information as soon as available and periodically issues an updated forecast.

Furthermore, Yue, Gong, Wang, Kan, Li & Ke (2014), proposed a cholera prediction model which focused based on the effect of climate factors in the estuary of Pearl River. In their research work, the climate data were collected daily at meteorological stations in Guangzhou and Shenzhen. Daily data were converted to monthly data. Parameter values such as water temperature coefficient, coefficient of cholera shifting in the regions, were determined by linear regression.

Cholera, as explained before, is among the deadliest disease, and proper prevention is required, the existing methods have tried to propose prediction models to predict cholera outbreaks. The previous models'

limitations include such as using fewer features for prediction, waiting until certain cases are reported before getting data, based on Rainfall, and based on the flowing speed of rivers, wind direction, and flood. In this research, more comprehensive cholera features would be used in predicting an outbreak before it occurs based on the existing secondary dataset using the Naïve Bayesian Classification technique. The proposed model has more features and is not dependent on certain events to occur before predicting any outbreak.

4. Materials and Methodologies

4. 1 Data Collection

The data used for the implementation of the classifiers based on naïve Bayes classification algorithms was retrieved from the Yobe State Ministry of health, the specifical dataset of five local governments namely, Gujba, Gulani, Damaturu, Potiskum, and Fune. In this research, a Secondary source was used as a source of data collection. Therefore, all medical records where cholera outbreaks occurred in previous years in Yobe State were considered and it was consists of many samples with different attributes and was converted as numeric input.

Furthermore, a total of 1443 datasets collected was used during this study, the data set was divided into training and testing data, which was further used in the development of the model using the Naïve Bayes Classification algorithm. After the data set was normalized, we were able to split the dataset into two sets; the training dataset and the testing set in a ratio of 80:20.

4.2 Data Pre-Processing

The sample structure of the dataset after collection from the Yobe State Ministry of Health consists of some missing data. Similarly, the dataset has been pre-processed and cleaned by fixing missing values, changing and mapping categorical values into binary, fixing data imbalance before it was further used to implement the model by dividing the data into a training set and testing set.
All the data set are found numeric type and class label otherwise referred to as target variable or expected of 0 and 1 only.

4.3 Implementation of the Proposed Framework

The techniques used for implementation was Naïve Bayes Classifier. The Bayesian classifier is adequate for calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have an improved probabilistic classifier. A Naive Bayes classifier regards that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable.

However, fitting of classifier from pre-processed data set requires implementing the relevant algorithm on the data set. The proposed framework is illustrated in figure 1 below. The model framework developed in figure 1, was simulated using a set of an untrained data set. The error generated was computed using the Null accuracy and confusion matrix.

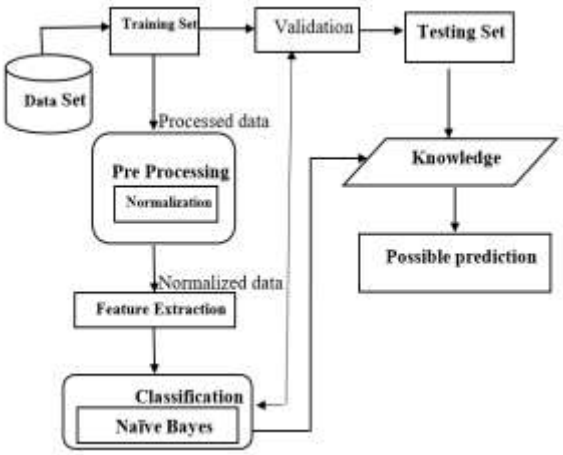

Figure 1: Proposed Framework for Cholera Prediction

5. Results

A total of 1443 datasets collected was used during this study, the data set was divided into training and testing data, which was further used in the development of the model using the Naïve Bayes Classification algorithm. Similarly, after the data set was normalized, we were able to split the dataset into two sets; the training dataset and the testing set in a ratio of 80:20.

5. 1 Model training and result prediction

After obtaining our model, we need to fit our model with the train set.

Table 1: Predicted Results of Training Dataset

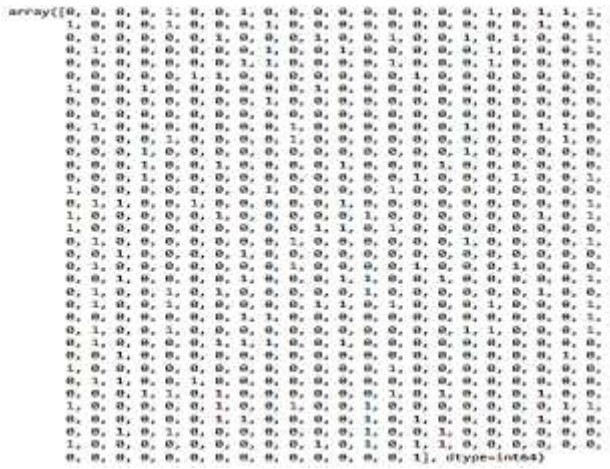

Table 1 above shows the result predicted after fitting the model with the training set. However, after training the model with the train set, it was found that the set accuracy score on the training set is 0.9949 as predicted by the Naïve Bayesian classification algorithm.

5.2 Model testing and result prediction

After we have successfully trained our model with the trained dataset, now we need to test our model using a test set and would be making predictions on it.

Table 2: Predicted Results on Test Set

Table 2 above shows that a successful result prediction on the test set and it was found that the model accuracy score on the test set is 0.9941. The result indicates the model designed has over a 99% accuracy rate compared to the normal traditional prediction techniques used which might give inaccurate result prediction.

5.3 Checking overfitting and underfitting

The training-set accuracy score on the set is 0.9949 while the test-set accuracy is 0.9941. These two values are quite comparable. Therefore, there is no sign of overfitting. When we compare our model accuracy with null accuracy, the model accuracy is 0.9941. But, we cannot say that our model is perfectly based on the above accuracy.

Therefore, another tool called Confusion matrix comes to our rescue. A confusion matrix was implemented for the error classifier identification and correction.

Table 3: Confusion Matrix for Error Classification on Test Set

Matrix	Description
True Positives (TP)	285
True Negatives (TN)	50
False Positives (FP)	1
False Negatives (FN)	1

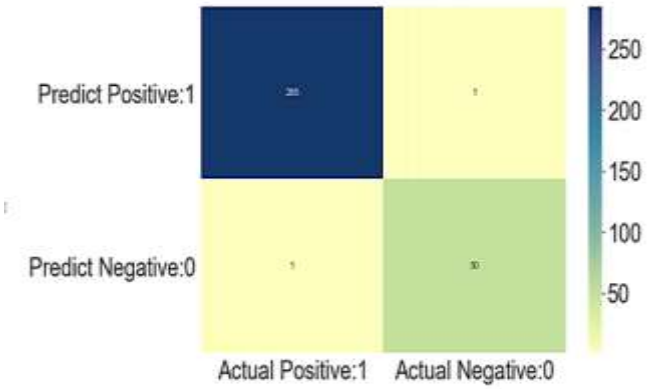

Figure 2: Confusion Matrix for Error Classification on Test Se

Table 3 and figure 2 above shows that we have True Positives (Actual Positive: 1 and Predict Positive: 1) as well as 285 True Negatives (Actual Negative:0 and Predict Negative:0) while 50 False Positives (Actual Negative:0 but Predict Positive:1) and 1 (Type I error) False Negatives (Actual Positive is 1 but Predict Negative:0) with 1 (Type II error).

5. 4 Classification Report

A classification report is another way to evaluate the classification model performance. It displays the precision, recall, F1, and support scores for the model.

Table 4: Classification Report for test set

	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	286
1	0.98	0.98	0.98	51
Accuracy			0.99	337
Macro Avg	0.99	0.99	0.99	337
Weighted Avg	0.99	0.99	0.99	337

Based on Table 4 above shows that, the proposed model classification accuracy is 0.9941. This indicates our model is reliable and can be applied in the health sector for predicting an accurate number of correct cases of cholera based on the available data.

In this study, the Bayesian Classification model was built to predict whether a person has a cholera infection or not. The model yields a very significant performance as indicated by the model accuracy which was found to be 0.9941 which is 99% accuracy.

6.0 Discussion

The training-set accuracy score is 0.9949 while the test-set accuracy to be 0.9941. These two values are quite comparable. Similarly, there is no sign of overfitting. We have compared the model accuracy score which is 0.9941 with a null accuracy score which is 0.8487. However, we can conclude that our Bayesian classifier model prediction of the class labels is significant.

The model is designed to predict a possible outbreak of cholera based on existing data collected from the case study. Experimentation was carried out on the datasets and the model has 99% accuracy, this has indicated the model prediction is significant. All the set objectives were achieved and a clear explanation was given in each section.

7. 0 Conclusion

In conclusion, this research aims to propose a cholera prediction Model using Feature Clustering Bayesian Classification that was successfully achieved based on the set objectives. The model was designed based on the conceptual framework outlined in Figure 1 above. The model was designed using Python programming language and also tested and trained using the collected dataset from the case study following a secondary source, the data were further divided into the testing set and training set. The result on the test set indicated high accuracy of 99%. This shows the model target was achieved and a more detailed explanation was given in Table 2 above.

7.1 Future Work

Based on the research aim, objectives, and analysis performed using the proposed model, the following are the recommendation for future work:

i. The model was designed and tested using a case study dataset, however, more dataset is required from different or multiple case studies to further enhance the accuracy of the model proposed.
ii. The model was proposed using Feature Clustering Bayesian Classification Machine learning algorithm, however, another ML algorithm can be used to see what the outcome of the accuracy rate is, and more benchmarking is needed.
iii. Mixed-method research should be applied to have a full understanding of the case study and also to reduce the chance of being biased.

7.2 Contribution to the Knowledge

This study advances the field of Machine Learning adoption to solving real-life problems, the previous research in predicting cholera outbreaks is a constraint to limited features, like (behavior, living situation), (wind direction, region, and speed of river flow), this feature cannot give a correct prediction as required, therefore, the proposed model consist of more features like (Leg cramps, House condition, Poor sanitation, Lethargy, Contaminated water, Fatigue, Fishy odour stool, Dry mucous membranes, Watery diarrhea, Irritability, Thirst, Muscle cramps, Rice water stool, Loss of skin elasticity, Rapid heart rate, Vomiting) used in the predicting the cholera outbreak. The previous work uses limited features for prediction which might predict with less accuracy. This model is a contribution to the knowledge as it indicates using more features in ML prediction gives the chance to have a more accurate prediction.

8. References

Leckebusch, G. C., & Abdussalam, A. F. (2015). Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Health & place*, *34*, 107-117.

Leo, J. (2020). A reference machine learning model for prediction of cholera epidemics based-on seasonal weather changes linkages in Tanzania (Doctoral dissertation, NM-AIST).

Mondiale de la Santé, O., & World Health Organization. (2017). Cholera vaccines: WHO position paper–August 2017–Vaccins anticholériques: Note de synthèse de l'OMS–août 2017. Weekly

Epidemiological Record= Relevé épidémiologique hebdomadaire, 92(34), 477-498.

Pasetto, D., Finger, F., Rinaldo, A., & Bertuzzo, E. (2017). Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. Advances in Wate Resources, 108, 345-356.

Reiner, R. C., King, A. A., Emch, M., Yunus, M., Faruque, A. S. G., & Pascual, M. (2012). Highly localized sensitivity to climate forcing drives endemic cholera i a megacity. *Proceedings of the National Academy of Sciences*, *109*(6), 2033-2036.

Salako, B. L., Ajayi, A. O., & Smith, S. I. (2021). Cholera in Nigeria: Epidemiology, Risk Factors, and Response-A Review. *Proceedings of the Nigerian Academy of Science*, *14*(1s).

Shirzad, E., Ataei, G., & Saadatfar, H. (2021). Applications of data mining in healthcare area: A survey. *Engineering and Applied Science Research*, *48*(3), 314-323.

Young, J. (2017). *Predicting Cholera Positive Cases in Haiti.* Haiti: ROCHESTER INSTITUTE OF TECHNOLOGY.

Yue, Y., Gong, J., Wang, D., Kan, B., Li, B., & Ke, C. (2014). *Influence of climate factors on Vibrio cholera dynamics in the Pearl River estuary, South China. World Journal of Microbiology and Biotechnology, 30(6), 1797–1808.* doi:10.1007/s11274-014-1604-5