Explorematics
Journal of Innovative
Engineering and Technology
Volume: 06 No: 01 | April -2025
ISSN (Online) 2636 – 590
ISSN (Print) 2636 - 591X

DEVELOPMENT OF A MACHINE LEARNING BASED MODEL FOR THE EARLY DETECTION OF CHOLERA

Bosede Olawale Joseph[*1], Lois Nwobodo[1], Chioma Oleka Violet[1]
1 Department of Computer Engineering, Enugu State University of Science and Technology
Author for correspondence: Bosede O.J; **E-mail:** walesbosede@gmail.com

Abstract - This study developed a machine learning-based model for detecting cholera, with a particular focus on using local data from the Nigeria Center for Disease Control (NCDC), Lagos state branch. The dataset, which contained patient data from patients who tested positive for cholera, was carefully used for training to ensure the model was relevant to the local context. To tackle the unique challenges posed by regional cholera outbreaks, the model was designed with a focus on ensuring its practical applicability in real-world settings. The subsequent phase of the study involves integrating the developed model into a web-based application to enable real-time cholera detection.`The model achieved impressive performance metrics, including a training accuracy of 0.9995, a validation accuracy of 0.9825, and a final test accuracy of 98.25%, demonstrating its strong capacity for generalization and robustness against overfitting. The model's robust performance and dependability make it ideal for real-world applications, particularly in the Nigerian healthcare system, even if it did not perform as well as other models. This model's incorporation into a web application for real-time cholera detection is a major advancement in improving medical practitioners' capacity to identify and treat cholera epidemics. The web-based application will offer a useful and effective tool that facilitates quicker and more precise diagnostics, reducing the spread and improving patient outcomes.

Keywords: Cholera Detection; Machine Learning; Nigeria Center for Disease Control; Lagos State

1. Introduction

Globally, 2.9million cholera cases have been estimated as annual reported cases of cholera (George et al., 2024). Cholera is an acute enteric diarrhea that is caused by Vibrio cholera bacteria (Kaisar et al., 2021), and over the years, it has presented a public health challenge, particularly in developing countries like Nigeria. This disease is characterized by severe watery diarrhea, which can rapidly lead to severe dehydration and death if not properly managed (Wiens et al., 2023). On the 9th of June, 2024, the Lagos state government announced a cholera outbreak, which has continued to claim lives to date and necessitates urgent attention (UNICEF, 2024). Cholera outbreak in Nigeria is not new, as Mart et al. (2024) reported that it has been a recurrent issue, with health and socio-economic implications. The motivation of this research stems from the need to develop a smart diagnostic system capable of early cholera detection and diagnosis using artificial intelligence techniques.

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

Currently, the odyssey of the Nigerian health care sector lacks the necessary resources, enough manpower, and technological capacity to manage the current cholera outbreak effectively, thus necessitating the need for solutions that will facilitate early detection and control of the cholera outbreak. In the scientific community, many studies have been presented in line to address this problem, considering various approaches such as biological, radiological, and Artificial Intelligence (AI) approaches, respectively (Weill et al., 2019; Zahra, 2019).

From the biological point of view, samples from the patient are collected and subject to several experiments and interpretations to identify cholera (Midani et al., 2020). In the case of the radiological approach, biological images of the test subject are collected and then analyzed to make findings, while in the case of A.I., data are collected that models the cholera disease and then applied to train a system that can help detect and manage cholera diseases very effectively (Ahmad et al. 2023).

AI, particularly machine learning algorithms and deep learning have continued to dominated studies on cholera detection, diagnosis and management, when compared with other approaches aforementioned due to several advantages such as high detection accuracy, ability to generate reliable model which facilitate fast diagnosis, easy to use, less resource intensive and reliable (Mart et al., 2024).

For instance, Ahmad et al. (2023) applied K-means clustering in developing an autoregressive algorithm for the detection of cholera and recorded 97% accuracy. Similarly, Onyijen et al. (2023) applied decision tree, random forest, and logistic regression algorithms on cholera datasets to develop a predictive model for forecasting the spread of cholera in West Africa. Similarly, Paul and Aurona (2023) trained Long Short Term Memory (LSTM), logistic regression, and decision trees, with data on cholera, and generated a model for its prediction. Although these studies have provided valuable insights into predicting cholera outbreaks, there is limited literature that specifically explores the use of machine learning for early-symptom-based cholera detection and patient diagnosis. This gap highlights the need for further research in this area. This study, therefore, proposed the development of an intelligent diagnostic system for the digital management of the new cholera using machine learning techniques.

2. Methodology

The methodology adopted for this study commenced with the systematic collection of cholera-related data, with emphasis on the prevailing Vibrio cholerae strains reported in Nigeria. The dataset was pre-processed to ensure accuracy and reliability before being applied to the development of a cholera detection model utilizing a machine learning approach, specifically an Artificial Neural Network (ANN). The developed model was subsequently incorporated into a clinical decision-support system designed to enhance the diagnosis of cholera within the Nigerian context. To evaluate its effectiveness, data from laboratory-confirmed cholera cases were used, and a comparative analysis was conducted to validate the model's diagnostic performance and accuracy.

2.1 Data collection

The data used in this study were collected from the Lagos State branch of the Nigeria Centre for Disease Control (NCDC). It consisted of documented cholera cases, including confirmed diagnoses, patient details, and associated epidemiological records collected during the defined study period. The data is made up of 27 cholera patients' medical records of both male and female genders who experience the symptoms of the disease, along with their various environmental factors. The dataset is later processed before being used to train the proposed machine learning algorithm for the early detection of cholera. The description of the data to be used is presented in Table 1, with their attributes and data types.

Table 1: Data Description

Attribute Name	Description	Data Type	Data Property
Age	Patient's Age	Integer	Non-negative optional null for missing data
Gender	Patient's Data	String	Values: Male or Female
Address	Residential location of the patient	String	Optional, non-unique
Severe Diarrhea	A large quantity of water in the stools	Boolean	True if positive and false if negative
Vomiting	Expulsion of stomach contents from the mouth	Boolean	True if present and false if absent
Severe dehydration	Experiences Dry mouth and sunken eyes	Boolean	True if present and false if absent
Thirst	Intense craving for water due to dehydration	Boolean	True if present and false if absent
Rapid weight loss	Significant and sudden loss of weight	Float	Measured in kilograms
Low blood pressure	Blood pressure below normal levels	Float	Measured in mmHg, basically <90/60
Rapid heartbeat rate	Noticeable increase in heartbeat rate	Integer	Measured in beats per minute (bpm), basically >100bpm
Muscle cramps	Painful contractions in the muscle due to electrolyte loss	Boolean	True if present and false if absent

Table 1 presents the attributes contained in the dataset along with their respective properties, which collectively capture indicators relevant to the presence of cholera in patients. This structured breakdown promotes clarity and uniformity in the recorded data, thereby supporting effective analysis and processing. Specifically, the Attribute Name column identifies each variable included in the dataset (e.g., Age, Gender, or Severe Diarrhea), the Description column provides a concise explanation of the variable's meaning (e.g., Patient's age in years), while the Data Type column indicates the nature of the data stored (e.g., Integer, Boolean, String), ensuring standardized recording and interpretation of values. Finally, Data Property outlines constraints or rules governing the data ("Non-negative", "Values: Male or Female", or "Measured in mmHg"), which enforce validity and usability for analysis. Together, these columns standardize data definitions, enabling accurate interpretation, seamless integration into databases, and reliable clinical or research applications.

2.2 Data processing

The acquired data goes through a processing phase, which is a stage that ensures clean, consistent, and prepared data for analysis by the proposed system. This stage involves the application of data cleaning at first, data cleaning handles missing values, removes outliers, and standardization of data formats using the Density-Based Spatial Clustering of Applications with Noise(DBSCAN) technique. The next stage of the processing involves the use of a correlation analysis approach for feature selection and SHapley Additive exPlanations (SHAP) for feature engineering, where the most relevant features are identified and created for predictive analysis through encoding of categorical variables and normalization of numerical features. Synthetic Minority Oversampling Technique (SMOTE) is applied for data balancing by recovering cases and randomly removing samples from the majority cases. Then, finally, the data integration stage is executed to merge external data sources to enrich the dataset and align data using unique identifiers.

2.3 The Proposed ANN Algorithm for Cholera Disease Detection

In this study, training of the ANN for the detection of cholera deals with several steps, which start with data processing, design of the model architecture, training of the model, and evaluation of the model performance. The architecture of the model used in this study contains 1 input node, then 2 densely connected layers with 64 and 32 neurons, respectively, and 1 neuron with a Sigmoid function in the output layer that classifies the presence of cholera in the data sample. The training process uses binary cross entropy to define the loss function, while, Adam optimizer is used for optimization of the model during training (Orji et al., 2024). To prevent overfitting during optimization, 60 epochs were applied during training, and Bayesian optimization was used for automatic hyper-parameter tuning using probabilistic models. The model was further implemented using the Python programming language before being tested to ascertain its performance evaluation.

2.3.1 The Classifications Model

The classification model presented in the study was the ANN-based model trained using the given dataset, which was split at a ratio of 70:30 for both the training and testing sets. The classification model utilizes the dataset obtained from the Lagos State branch of the Nigeria Centre for Disease Control (NCDC) as input. Before model training, the data undergoes a series of pre-processing steps, including cleaning, feature selection and engineering, normalization, and the treatment of class imbalance. The data is further integrated before being fed into the trained ANN model. The trained model employed the Rectified Linear Unit (ReLU) as the activation function across the two fully connected hidden layers, while the Adam optimization algorithm was applied to refine the training process, which was executed over 60 epochs. The output of the classification model presents the result of implementation by classifying the presence of cholera or its absence. The model of the proposed system is presented in Figure 1, representing the steps and components of the model

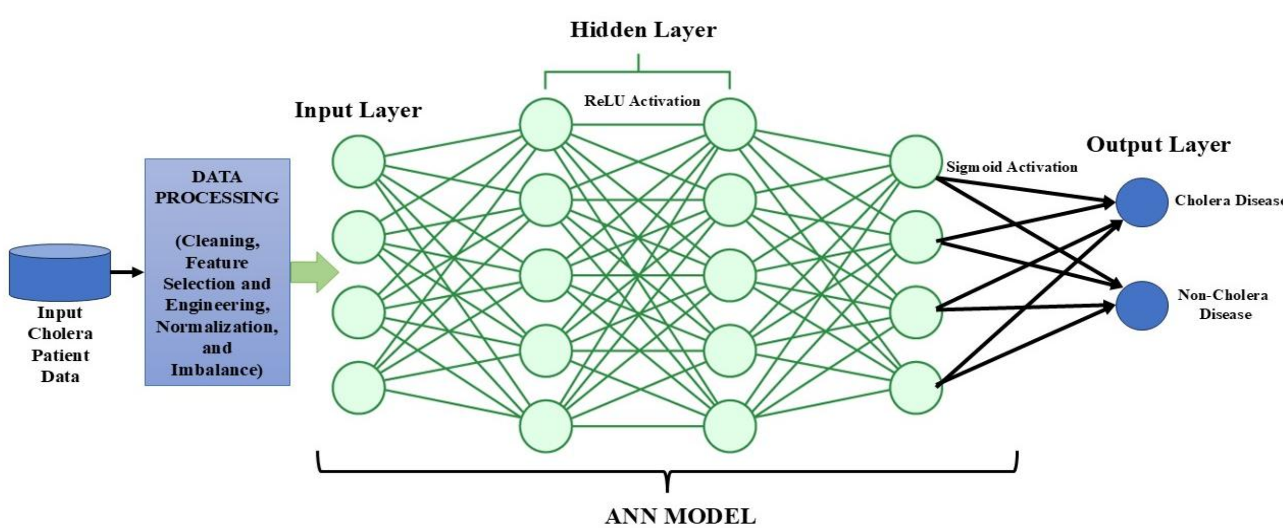

Figure 1: Model of the Proposed Cholera Detection System

2.3.2 Flowchart of the Proposed Cholera Diagnostic System

The diagnostic system presents the use of data from the Nigeria Center for Disease Control, Lagos State chapter, to populate the model for the early classification and diagnosis of cholera disease. Then, according to the flowchart in Figure 2, the diagnostic system further recommends steps to manage the cholera disease in the patient.

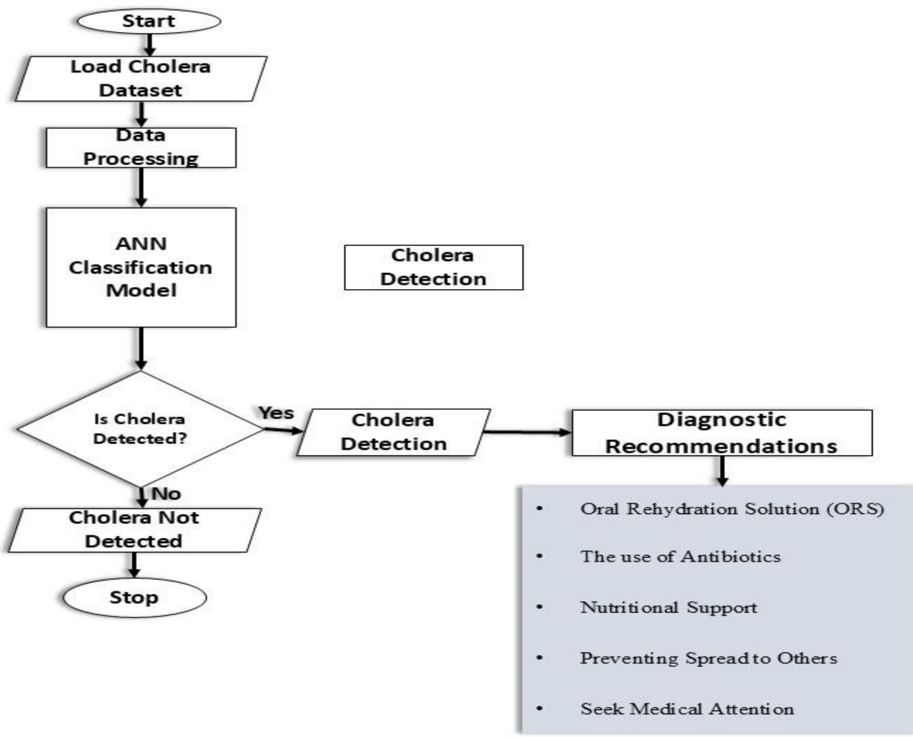

Figure 2: System Flowchart of the Cholera Diagnostic System

In Figure 2, the flowchart presents the sequential steps taken by the system for the early diagnosis and management of cholera. In the system, after the ANN model has undergone training, data of a patient is collected and fed into the classification model, which decides whether the data is infected with cholera or not. When cholera is not detected, the system prompts the user with a report of the absence of cholera. When cholera is detected, the system provides recommended diagnostic procedures to guide the management of the identified case.

2.3.4 The Diagnostic Model of the System

This section of the model is responsible for making a diagnostic decision on the classification results obtained from the classification model. The diagnosis of cholera disease involves clinical evaluation, laboratory tests, and rapid diagnostic tools. With the advancement of medical AI, deep learning, and ANN models, the technology can be applied for the diagnosis of cholera based on clinical symptoms and laboratory data. The diagnostic model presented in this study uses an ANN model for the classification of the cholera status of a patient as either a cholera or a non-cholera patient.

On the detection of cholera in a patient, immediate response is recommended for the prevention of severe situations like intense dehydration, complications, and death. Effective management of the condition requires interventions such as rehydration therapy, administration of antibiotics, nutritional support, and the implementation of preventive measures to curb further transmission. Table 2 outlines the essential steps patients should follow to ensure proper management of the disease.

Table 2: Recommended steps for the management of cholera detection (Source: Nigeria Center for Disease Control, Lagos State)

Steps	Recommended Steps After Diagnosis
1	**Oral Rehydration Solution (ORS)** As recommended by the World Health Organization (WHO), the ORS solution, which contains sodium, potassium, chloride, glucose, and water, should be used to restore fluids into the body system. The ORS should be taken in small and frequent sips **Intravenous (IV) Fluids** In the case of severe dehydration, IV Ringer's lactate solution should be administered to the patient in a hospital setting. When electrolyte levelsare monitored to prevent complications.
2	**The use of Antibiotics** To reduce the severity and duration of diarrhea, Doxycycline antibiotics should be taken in a single dose. Then, Azithromycin is an alternative that can be taken by children or pregnant women, and Ciprofloxacin or Erythromycin are other bacterial-resistant options that can be taken. However, overuse of antibiotics can lead to drug resistance
3	**Nutritional Support** The consumption of soft, easily digestible foods like rice, boiled potatoes, bananas, soups, and porridge is recommended. High potassium foods like oranges and coconut water can also be taken to replenish electrolytes. Cooked eggs, fish or lentils can be used as protein sources. However, it is recommended that the patient avoid spicy, fatty, and dairy-rich foods until digestion improves.
4	**Preventing Spread to Others** The patient must take proactive actions to prevent the spread of the disease to other people by hand hygiene practices, safe drinking water adoption, proper disposal of wastes, food safety by eating only cooked foods, and avoiding raw seafood and street food in outbreak areas
5	**Seek Medical Attention** It is recommended that the patient should immediately return to the hospital if they experience extreme weakness, confusion, or no urination, persistent vomiting or inability to keep ORS down, severe diarrhoea despite treatment, and signs of shock due to low blood pressure and rapid heart rate.

Cholera can be fatal if left untreated, but with rapid rehydration therapy, supportive care, and preventive measures, patients can recover fully within a few days. Proper hygiene and vaccination play a key role in preventing future outbreaks.

2.6 System Implementation

The implementation of a Cholera Diagnosis System using Python and Artificial Neural Networks (ANNs) involves multiple key steps. In order to handle missing values, normalise features, and encode categorical variables, the dataset must first be imported and pre-processed. Techniques for feature selection aid in determining which symptoms are most pertinent to a diagnosis. The Artificial Neural Network (ANN) model was implemented using TensorFlow/Keras and comprised an input layer, multiple hidden layers utilizing the ReLU activation function, and an output layer with a sigmoid activation function to perform binary classification. The Adam optimiser and binary cross-entropy loss function are used to train the model using pre-processed data, and accuracy is the basis for assessment. Following training, the model uses patient symptoms to predict the presence of cholera, and its performance is evaluated using visualisation approaches. Further improvements to the system can be achieved through hyperparameter optimization, deployment using Flask or FastAPI frameworks, and real-time integration with healthcare platforms to enhance disease identification and management.

Results of the ANN Training

The study's findings were taken from a Google Colab notebook. Training and validation metrics are shown for epochs 1 through 50 of the machine learning model that is being trained on the notebook. The time spent on each step, training accuracy, training loss, validation accuracy, and validation loss are some of these variables. The model was successfully learning from the training data, as evidenced by the implementation's high training accuracy of 0.9995 and low training loss of 0.0308. Similarly, Figure 3 shows that the system obtained a reasonably low validation loss of 0.0709 and a high validation accuracy of 0.9825, indicating that the model is generalising effectively to new, unknown data.

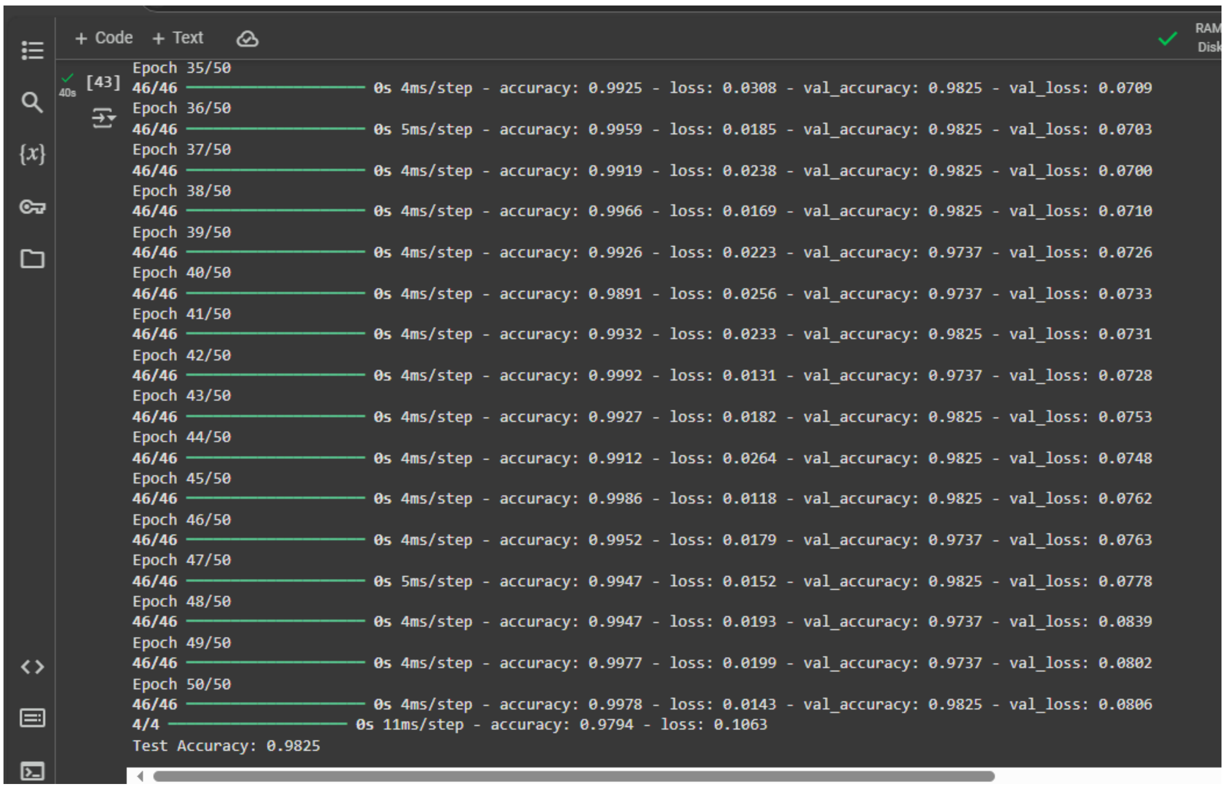

Figure 3: Training Epochs of the ANN model

Furthermore, the robust performance of the model is confirmed by the final test accuracy of 0.9825, which is shown at the bottom of the output. These indicators are essential for assessing the machine learning model's dependability and efficacy. As illuminating proof of the model's development, the screenshot in Figure 4 shows how well the model can compress, reconstruct, and forecast data, underscoring its potential use in cholera-related or other data-driven real-world applications. The plot of the ANN model's performance accuracy across 50 epochs is shown in Figure 4.

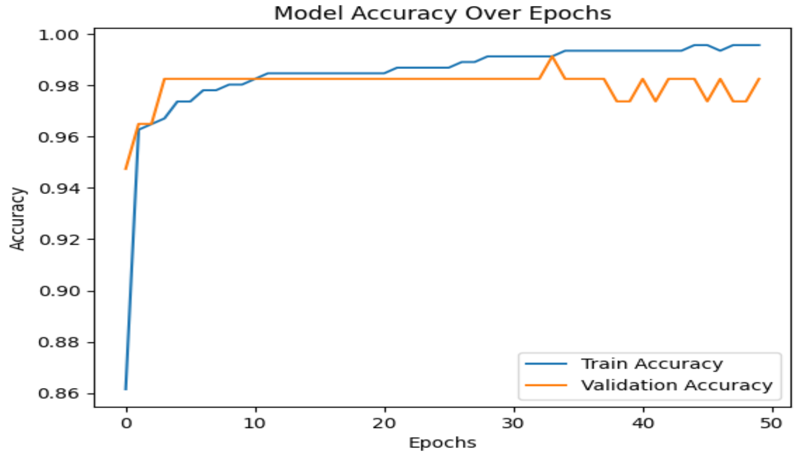

Figure 4: ANN Model Performance Accuracy

The notebook displays the outcome in Figure 5, which describes the training process of a machine learning model across several epochs. For epochs 1 through 50, it provides a list of performance indicators, including training accuracy, training loss, validation accuracy, and validation loss. Notably, the model has successfully learnt from the training data, as evidenced by the high training accuracy values of 0.9995 and low training losses of 0.0308. With validation accuracy values constantly high at 0.9825 and validation losses comparatively low at 0.0709, the validation metrics further demonstrate the model's capacity to generalise to new data.

The study's final test accuracy of 0.9825 highlights the model's strong performance. These findings show that the model is capable of producing precise predictions on unseen data and show that it is not overfitting. The image's snapshot successfully illustrates the model's effectiveness in data learning, compression, and prediction, indicating its possible use in cholera-related real-world activities or other data-driven domains.

4.2 System Integration

The developed ANN model for the detection of cholera disease in patients was further integrated as a web-based application. The application was developed using Web development toolboxes such as Hypertext Markup Language (HTML), Cascading Stylesheet (CSS), and JavaScript. The HTML case of the study was built to handle the layout of the application for easier user interaction, while CSS was used to improve the beauty of the layout, adding colors, fonts, and alignments to page layouts. The interactive functionalities of the web platform were implemented using JavaScript, enabling the buttons to respond dynamically and perform user-driven actions. JavaScript equally triggers the analytical functions of the platform to classify whether cholera is diagnosed or not. The homepage layout of the platform is presented in Figure 5.

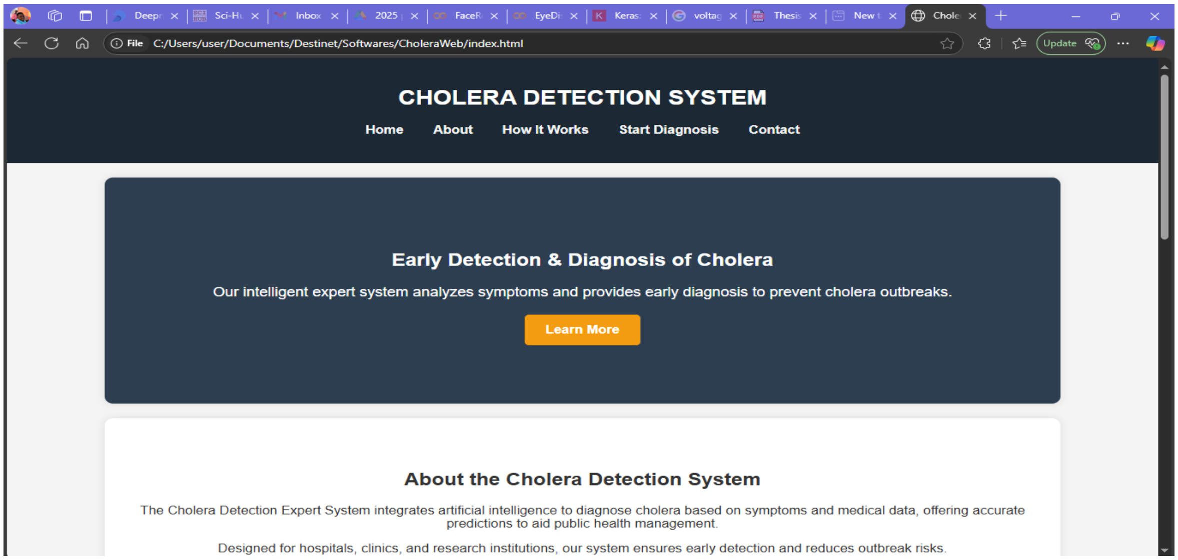

Figure 5: System Homepage

In Figure 5, the display of the system landing page is presented. The page shows the title and navigation labels of the platform; clicking on the labels takes the user directly to access the information required from the system. Figure 6 presents the layout where the patient can key in their necessary information for diagnosis of cholera.

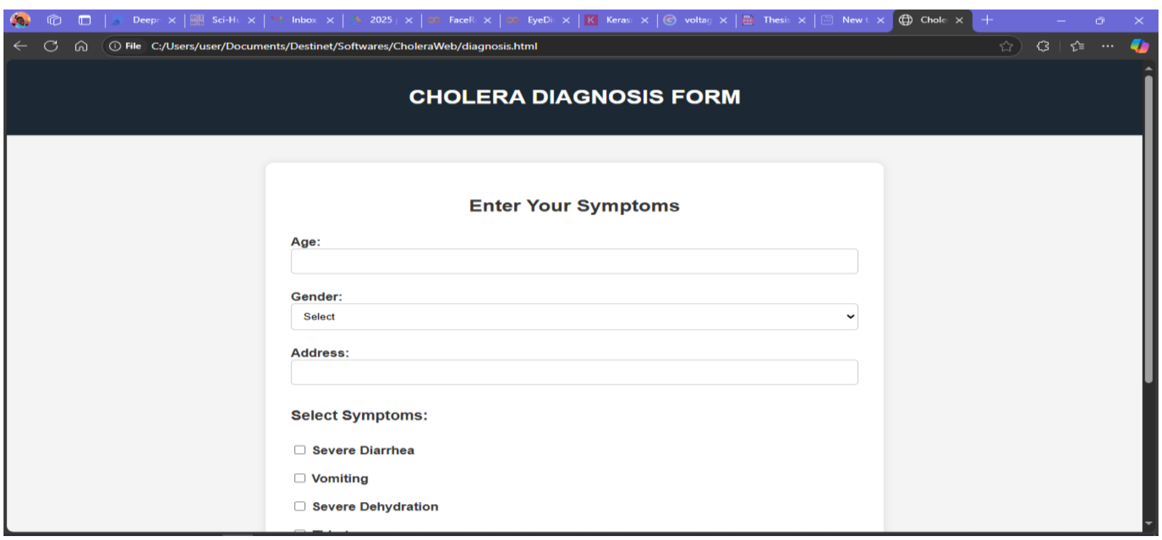

Figure 6: Cholera Diagnosis Page

For the diagnosis of the disease, the patient is required to input their information and identify symptoms they experience by checking the necessary checkboxes. The system uses the information keyed in by the user to determine the health status of the patient. When the patient is diagnosed with cholera disease, the system takes the user to a recommendation page, where necessary actions are advised for managing the severity of the health situation. This is done to mitigate complex health situations that may arise from late cholera diagnosis. Figures7 and 8present the outcome of cholera diagnosis

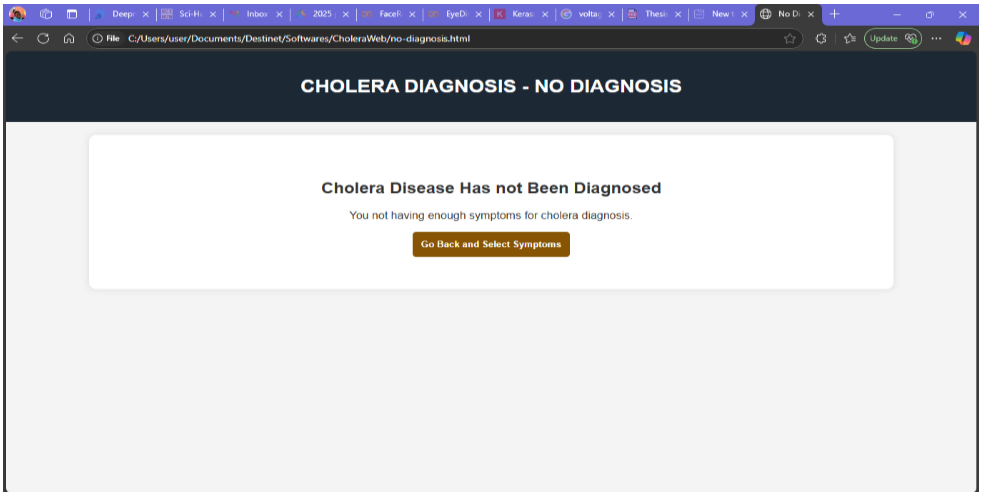

Figure 7: No Cholera Diagnosed Page

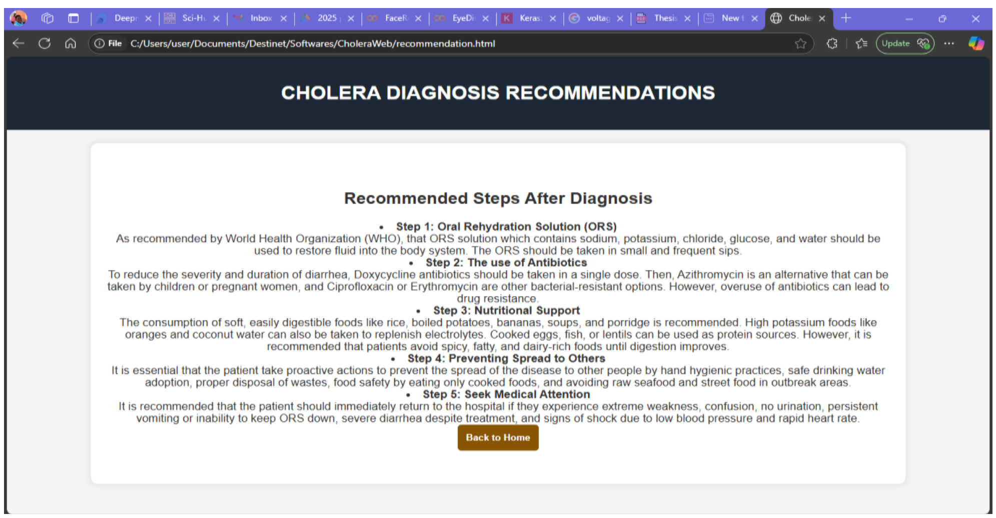

Figure 8: Recommendations after Cholera was diagnosed

In Figures 7 and 8, the resultant outcomes of the system are shown when the patient is either diagnosed with cholera or not. Figure 8 reports to the user how his/her system is free from the disease and gives a button to go back to the previous page. Figure 9 shows the page where recommendations are made to the user in order to manage the health situation when cholera is detected.

3. Conclusion

In conclusion, our work used a local dataset from the Lagos state branch of the Nigeria Centre for Disease Control to effectively construct a machine learning-based model for cholera diagnosis. The model's robust performance and dependability make it ideal for real-world applications, particularly in the Nigerian healthcare system, even if it did not perform as well as other models. The model's efficacy in correctly detecting cholera is ensured by its utilisation of local data, which is essential for enhancing public health interventions in the area.

This model's incorporation into a web application for real-time cholera detection is a major advancement in improving medical practitioners' capacity to identify and treat cholera epidemics. The web-based application will offer a useful and effective tool that facilitates quicker and more precise diagnostics, reducing the disease's spread and improving patient outcomes. All things considered, the study shows how machine learning and localised data may be used to solve public health issues, emphasising the value of regionally unique solutions. The study helps create more dependable, easily available, and efficient healthcare solutions by fusing cutting-edge technology with context-specific data, opening the door for better illness diagnosis and prevention initiatives in Nigeria.

References

Ahmad, H. A., & Prasad, R. (2023). Time series analysis and forecasting of cholera disease using discrete wavelet transform and seasonal autoregressive integrated moving average model. Scientific African, 20, e01652, 1–12.

George, C. M., Namunesha, A., Felicien, W., Endres, K., Luo, W., Bisimwa, L., Williams, C., et al., (2024). Evaluation of a rapid diagnostic test for the detection of Vibrio cholerae O1 in the Democratic Republic of the Congo: Preventive intervention for cholera for 7 days (PICHA7 program). Tropical Medicine & International Health, 29(7), 594–598. https://doi.org/10.1111/tmi.13998

Kaisar, M. H., Bhuiyan, M. S., Akter, A., et al. (2021). Vibrio cholerae sialidase-specific immune responses are associated with protection against cholera. mSphere, 6(1), e01232–20.

Mart, J., Olatuja, G., & Oni, T. (2024). Using artificial intelligence and machine learning to curtail cholera outbreaks in Nigeria. https://doi.org/10.13140/RG.2.2.22338.39367/1

Midani, F. S., Weil, A. A., Chowdhury, F., et al. (2020). Human gut microbiota predicts susceptibility to Vibrio cholerae infection. Journal of Infectious Diseases, 218(4), 645–653.

Oluwadare, A. (2020). Cholera outbreaks in Nigeria: An overview. Journal of Public Health and Epidemiology, 6(2), 25–30.

Onyijen, O. H., Olaitan, E. O., Olayinka, T. C., & Oyelola, S. (2023). Data-driven ML techniques for the prediction of cholera outbreaks in West Africa. International Journal of Applied and Natural Sciences. http://bluemarkpublishers.com/index.php/IJANS

Orji, E. Z., Haydar, A., & Erşan, İ. (2024). Inception-Transformer: End-To-End Printed Mathematical Expression Recognition Using Convolutional and Transformer Networks. Nanotechnology Perceptions, 20(6), 2672–2690. https://doi.org/10.62441/NANO-NTP.VI.3274

Osuolale, O., & Okoh, M. (2016). Review of cholera epidemiology in Nigeria. Water, Air, & Soil Pollution, 227(11), 1–15.

Paul, J., & Aurona, G. (2023). ML-based NLP for emotion classification on a Cholera X dataset. Center for AI Research (CAIR), South Africa.

UNICEF. (2024). Nigeria Sitrep: Lagos cholera outbreak – 14 June 2024. https://www.unicef.org/media/158376/file/Nigeria-Humanitarian-SitRep-Lagos-Cholera-Outbreak-14-June-2024

Weill, F.-X., et al. (2019). Genomic insights into the 2016–2017 cholera epidemic in Yemen. Nature, 565, 230–233. https://doi.org/10.1038/s41586-018-0818-3

Wiens, K. E., Iyer, A. S., Bhuiyan, T. R., Lu, L. L., Cizmeci, D., Gorman, M. J., Yuan, D., Becker, R. L., Ryan, E. T., Calderwood, S. B., LaRocque, R. C., Chowdhury, F., Khan, A. I., Levine, M. M., Chen, W. H., Charles, R. C., Azman, A. S., Qadri, F., Alter, G., & Harris, J. B. (2023). Predicting Vibrio cholerae infection and symptomatic disease: A systems serology study. The Lancet Microbe, 4(4), e228–e235. https://doi.org/10.1016/S2666-5247(22)00391-3

World Health Organization (WHO). (2024). Cholera in the WHO African region: Weekly regional cholera bulletin. https://www.afro.who.int/health-topics/disease-outbreaks/cholera-who-afr

Zahra, P., Tafazzoli-S., M., Nejadgholi, I., Mansourian, A., & Rahbar, M. (2016). Model of cholera forecasting using artificial neural network in Chabahar City, Iran. International Journal of Enteric Pathogens, 4, e31445. https://doi.org/10.17795/ijep31445