ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 9 (Issue 2) (2025) Pages 38-45

A Context-Aware End-to-End Predictive Analytics Architecture for Cholera Early Warning and Medical Resource Allocation Using Machine Learning

Kabwenda Moonga
School of Computing, Technology and Applied Sciences
ZCAS University
Lusaka, Zambia
kabwendamoonga@gmail.com

Aaron Zimba
School of Computing, Technology and Applied Sciences
ZCAS University
Lusaka, Zambia
aaron.zimba@zcasu.edu.zm

***Abstract*—This paper details the architecture, development, and assessment of a comprehensive predictive analytics platform designed to forecast cholera outbreaks and optimize the allocation of medical resources within Zambia's public health system. Conventional cholera management in the region is predominantly reactive, resulting in operational delays and suboptimal resource deployment. This research confronts this issue by creating a localized, proactive decision-support tool. The system utilizes a hybrid modeling approach, combining supervised learning algorithms (Logistic Regression, Random Forest, XGBoost) with a linear programming (LP) model for resource optimization. A comparative analysis was performed using a synthetic dataset from 2017-2024 that mirrors Zambia's epidemiological trends. The XGBoost model yielded the most effective performance for an early warning system, attaining an accuracy of 84.69%, a flawless recall of 1.0, and an AUC-ROC score of 0.9361. Conversely, the Random Forest model provided perfect precision (1.0) but with a minimal recall of 0.125, underscoring significant performance trade-offs. The resulting prototype, which includes an interactive Streamlit dashboard, effectively transforms predictive outputs into actionable resource allocation strategies, offering a scalable and data-driven solution for epidemic preparedness in resource-limited settings.**

Keywords*—*Predictive Analytics, Machine Learning, Cholera Forecasting, Resource Optimization, XGBoost, Decision Support System, Zambia, Public Health Informatics

I. INTRODUCTION

Like many nations in sub-Saharan Africa, Zambia perpetually contends with the challenge of managing communicable disease outbreaks, among which cholera remains a formidable public health concern. These outbreaks, which intensify during rainy seasons due to compromised sanitation, exert immense pressure on the national healthcare infrastructure. The 2017-2018 outbreak led to over 5,900 cases and 114 fatalities, while a more recent event from October 2023 to February 2024 saw over 19,700 cases, highlighting the problem's severity and persistence [4].

Historically, crisis management strategies in Zambia have been largely reactive. Interventions such as the emergency deployment of Water, Sanitation, and Hygiene (WASH) resources, public health campaigns, and medical supply distribution typically begin only after a substantial number of cases have been confirmed. This reactive approach frequently causes delays in resource mobilization, introduces inefficiencies in healthcare services, and ultimately contributes to higher rates of morbidity and mortality [5]. The fundamental issue is the absence of a data-driven, localized early warning system capable of forecasting high-risk zones to guide proactive interventions.

Globally, machine learning (ML) has demonstrated significant potential to transition public health from a reactive to a proactive paradigm [8]. Through the analysis of intricate patterns within historical health, climate, and demographic data, ML models can predict outbreaks with an accuracy that facilitates timely preparation. However, a majority of existing models are not adapted to Zambia's specific epidemiological context and often lack a direct pathway to convert predictions into concrete

operational plans [19]. This paper aims to bridge these gaps by introducing a holistic, end-to-end system engineered specifically for the Zambian environment.

II. RELATED WORKS

The use of machine learning for infectious disease forecasting is a mature research area focused on enabling proactive public health responses. While numerous studies have confirmed the potential of various algorithms, significant challenges persist regarding localized adaptation and the operationalization of predictive outputs. A summary of key related works is presented in TABLE I.

Several studies have investigated cholera prediction in environments comparable to Zambia. Ibrahim et al. [24] employed Random Forest and XGBoost in Nigeria, achieving 87% accuracy but underscoring the need for a robust data infrastructure. In Malawi, Ghosha et al. [23] applied time-series models like LSTM, reaching 83% accuracy for short-term forecasts but noted limitations in scalability. Similarly, Leo [25] used Decision Trees in Tanzania based on climate data, but the solution lacked real-time deployment capabilities. Within Zambia, Musakuzi [26] completed a feasibility study using hybrid ML models in Lusaka Province, which achieved 88% accuracy but was not advanced to a large-scale, deployable system.

A review of the literature reveals three critical gaps. First, a **lack of contextualization** exists, as many models are trained on generic datasets and fail to incorporate Zambia's unique epidemiological and socioeconomic factors, thus limiting their local relevance [20]. Second is an **actionability gap**, where most research concludes at prediction without guiding subsequent operational steps, creating a need for systems that integrate forecasting with resource allocation [87]. Finally, an **implementation gap** is evident, as many studies remain theoretical without being developed into user-friendly, end-to-end systems for non-technical health professionals.

This research directly confronts these shortcomings by delivering a system that is: (1) **Context-Specific**, using a hybrid framework trained on parameters relevant to Zambia; (2) **Actionable**, by integrating predictions with an LP model for resource optimization; and (3) **Operational**, providing a functional prototype with an interactive dashboard for practical use.

TABLE I. SUMMARY OF RELATED WORKS.

Study	Algorithms Used	Dataset	Accuracy	Key Findings / Limitations
Mbunge & Batani [20]	CNN, RNN	Multi-country health datasets	0.85	Shows ML potential in healthcare but lacks regional adaptation.
Carter [22]	SVM, Decision Trees, Neural Networks	Multiple healthcare datasets	0.9	Broad overview, not case-specific.
Ghosha et al. [23]	Time-series, LSTM	Malawi cholera data	0.83	Effective for short-term forecasts, limited scalability.
Ibrahim et al. [24]	Random Forest, XGBoost	Nigerian cholera data	0.87	Strong detection but requires robust data infrastructure.
Leo [25]	Decision Trees, Bayesian Networks	Seasonal climate & health data	0.8	Climate-based prediction but lacks real-time deployment.
Musakuzi [26]	Hybrid ML models	Lusaka health records	0.88	Feasibility shown, lacks scale-up implementation.
Onyijen & Tosin [27]	Logistic Regression, KNN	West African epidemic data	0.82	Emphasizes regional data but low adaptability.
Urukadle [28]	Deep Learning, CNN	Hospital patient data	0.91	High predictive power, computationally intensive.
Mudenda & Mohamed [29]	Random Forest, XGBoost	Global pandemic datasets	0.88	Good detection, but needs real-time adaptation.

III. METHODOLOGY

This work was framed using the Design Science Research (DSR) paradigm, which centers on creating and assessing an artifact to address a practical problem [30]. The artifact, our Cholera Outbreak Prediction and Optimization System, was developed following the DSR cycle as depicted in Fig. 1.

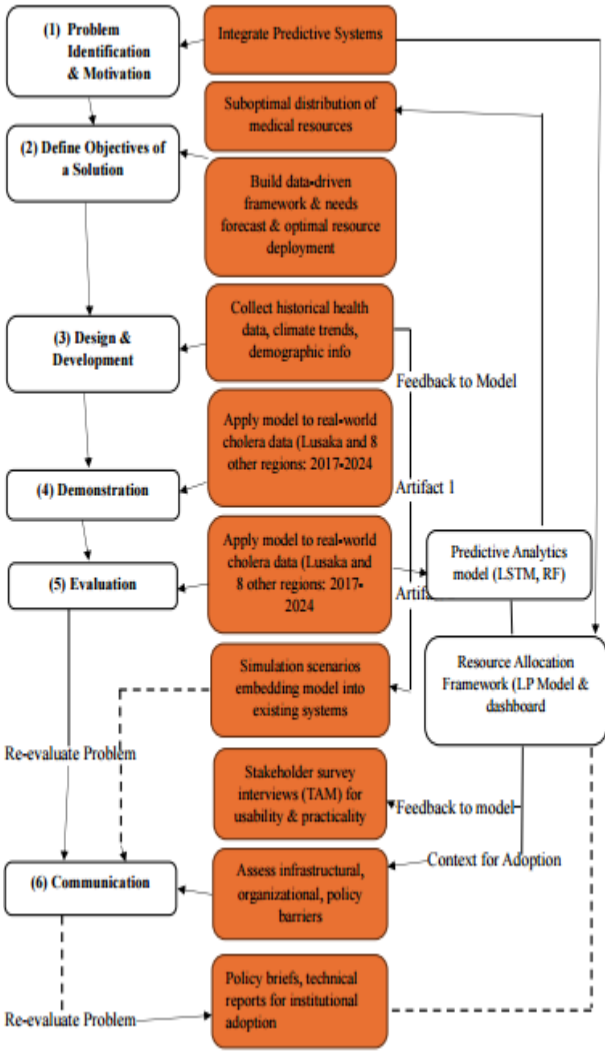


Figure 1. DSR Process for Cholera Outbreak Management in Zambia.

A. Data Collection and Pre-processing

The model was engineered to use a variety of datasets that reflect cholera's primary drivers. Due to the absence of ethical clearance for real-world patient data, a detailed synthetic dataset for the 2017-2024 period was created to emulate actual data patterns. This included health data (weekly cases), climate data (rainfall, temperature), and demographic data (population density, sanitation access).

The pre-processing pipeline, executed via the data_preprocessing.py script, involved data cleaning with median imputation for missing values, Min-Max scaling for normalization, and feature engineering. New predictive variables, such as lagged features and rolling window statistics (e.g., four-week average rainfall), were created to capture vital temporal and environmental context.

B. Predictive Modeling

A hybrid, multi-model strategy was employed to determine the optimal algorithm for the binary classification task of predicting an outbreak. The evaluated supervised learning models included:

- **Logistic Regression (LR):** A baseline for interpretability.
- **Random Forest (RF):** An ensemble method for handling non-linear relationships.
- **XGBoost (Extreme Gradient Boosting):** An efficient gradient-boosting algorithm suitable for imbalanced datasets.
- **Long Short-Term Memory (LSTM):** A RNN to capture time-series dependencies.

Model training and evaluation were managed by the model_training.py and evaluate.py scripts. The dataset was partitioned into training (80%), validation (10%), and testing (10%) sets. Performance was assessed using standard classification metrics (Accuracy, Precision, Recall, F1-Score, AUC-ROC) at optimized decision thresholds.

Prediction Pseudo-Algorithm

To operationalize outbreak prediction, a structured algorithm was developed to standardize model inference and decision-making.

PROCEDURE Generate_Outbreak_Prediction

INPUT:

- *New_Data_File*: Latest raw data (CSV) containing health, climate, and demographic indicators per district.
- *Selected_Model_Type*: The pre-trained model for inference (e.g., 'XGBoost', 'Random Forest').

OUTPUT:

- *Prediction_Result*: Outbreak/No Outbreak classification for each district.
- *Risk_Score*: Outbreak probability (0–1) per district.

BEGIN

1. Initialize System and Load Artifacts
    - Load ML_Model (e.g., xgboost_model.joblib).
    - Load Feature_Scaler (minmax_scaler.joblib).
    - Load Expected_Features list.
    - Load Decision_Threshold.
2. Ingest and Preprocess New Data
    - Read Raw_Input_Data from *New_Data_File*.
    - Apply feature engineering (lagged & rolling features).
    - Validate against Expected_Features and reorder columns.
    - Apply Feature_Scaler to produce Prepared_Data.
3. Perform Model Inference
    - Pass Prepared_Data into ML_Model.
    - Generate raw probability output as Risk_Score.
4. Apply Decision Threshold
    - For each district:
        - If Risk_Score ≥ Decision_Threshold → Prediction_Result = Outbreak.
        - Else → Prediction_Result = No Outbreak.
5. Return and Display Results
    - Return Prediction_Result and Risk_Score.
    - Visualize in dashboard (tabular + graphical).

END PROCEDURE

C. Resource Allocation Framework

A significant innovation of this research is the fusion of predictive outputs with a resource optimization module. A Linear Programming (LP) model, implemented in resource_optimization.py, was used for this purpose. The model's objective is to maximize the total risk-weighted impact of resource deployment, formulated as:

Maximize $Z = \Sigma (R_i \cdot c_{ij} \cdot x_{ij})$

where R_i is the predicted risk score for region *i*, c_{ij} is the effectiveness of resource *j* in region *i*, and x_{ij} is the quantity of resource *j* allocated. The model is constrained by the total availability of each resource. This framework translates the "where and when" of a forecast into a "what to do" recommendation.

D. Model Evaluation

Predictive models were evaluated using metrics derived from the confusion matrix (TP,TN,FP,FN)(TP, TN, FP, FN)(TP,TN,FP,FN). The selected metrics include:

- **Accuracy:**

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [4]$$

- **Precision:** proportion of predicted outbreaks that were actual outbreaks.

$$Precision = \frac{TP}{TP+FP} \quad [5]$$

- **Recall (Sensitivity):** proportion of actual outbreaks correctly identified.

$$Recall = \frac{TP}{TP+FN} \quad [6]$$

- **F1-Score:** harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot (Precion \, . Recall)}{Precision+Recall} \quad [7]$$

- **AUC-ROC:** area under the Receiver Operating Characteristic curve, reflecting the model's discriminatory power across thresholds.

IV. EXPERIMENT/SETUP

The prototype was developed as a modular Python application with distinct components for offline

batch processing and online real-time interaction, as shown in the architecture diagram in Figure. 2.

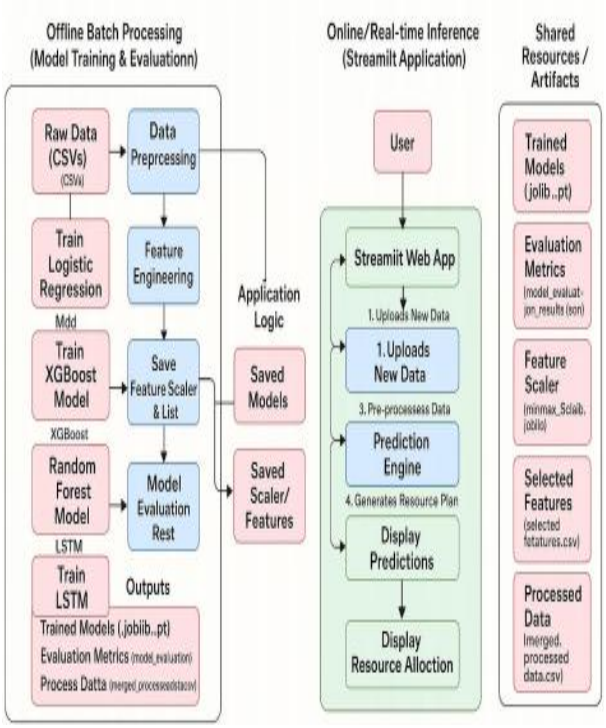


Figure 2. System Architecture Diagram showing Offline and Online Modes.

A. System Modules

1. **Data Ingestion Layer:** This layer, managed by data_preprocessing.py, is responsible for loading, cleaning, and transforming raw data from multiple CSV sources into a single, unified dataset ready for the ML Engine.
2. **Machine Learning (ML) Engine:** This core predictive component contains scripts for feature engineering (feature_engineering.py), model training (model_training.py), and prediction (predict.py). Trained models are serialized and saved (e.g., as .joblib files) for reuse.
3. **Optimization Module:** Implemented in resource_optimization.py, this module takes the prediction outputs from the ML Engine and, using the LP model, generates a practical resource allocation plan, which is exported as a CSV file.
4. **Interactive Dashboard:** A user-friendly graphical interface built with Streamlit (app.py). It allows non-technical users to upload new data, run predictions using pre-trained models, and visualize both the outbreak risk and the recommended resource allocation plan in real-time.

B. Operational Modes

The system features two operational modes to separate computationally intensive training from real-time decision-making, as illustrated in Fig. 3.

- **Offline Batch Processing:** Initiated via main.py, this mode is used for training or retraining models on updated historical data. It runs the full data ingestion, preprocessing, training, and evaluation pipeline, saving all resulting artifacts (models, scalers, evaluation metrics) for future use.
- **Online Real-Time Interaction:** Activated via app.py, this mode launches the Streamlit dashboard. It loads the pre-trained models and artifacts saved during the offline process. This allows health officials to upload new, current data (e.g., the latest week's rainfall and case numbers) and receive immediate predictions and resource recommendations without the delay of retraining.

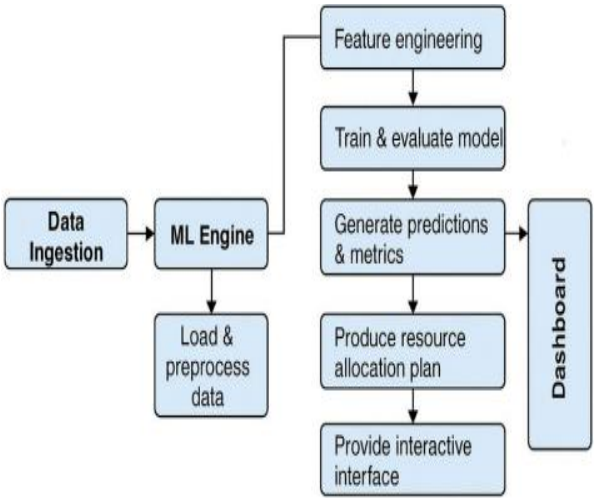


Figure 3. Low-level Diagram of Main Prototype Functions.

This decoupled architecture ensures that the intensive task of model training does not interfere with the system's real-time usability, making it a practical tool for dynamic public health environments.

V. RESULTS AND DISCUSSION

The models were rigorously evaluated on the test set at optimized decision thresholds. The comparative performance is summarized in TABLE II.

TABLE II. PERFORMANCE COMPARISON OF OUR MODELS VS. RELATED STUDIES

Study/System	Algorithms Used	Accuracy	Precision	Recall	AUC-ROC	F1-Score	Key Observations
Our System – Logistic Regression	Logistic Regression	0.7755	0.0625	0.125	0.4625	0.0833	Moderate accuracy; low precision an limited ability to detect outbreaks at optimized threshold.
Our System – XGBoost	XGBoost	0.8469	0.3478	1.0	0.9361	0.5161	High recall and AUC-ROC; strong at detecting all outbreaks, but some false positives remain.
Our System – Random Forest	Random Forest	0.9286	1.0	0.125	0.8264	0.2222	Very high precision but low recall; d few outbreaks, but when it does, it is always correct.
Our System – LSTM	LSTM (PyTorch)	0.2614	0.0857	0.8571	0.4744	0.1558	Low accuracy, but high recall; good at catching outbreaks, but with false positives.
Mbunge & Batani [20]	CNN, RNN	0.85	0.82	0.8	0.87	N/A	Strong for general health prediction, regional specificity.
Zoe Carter [22]	SVM, DT, Neural Networks	0.9	0.88	0.85	0.91	N/A	Broad applicability, but lacks cholera use case.
Ghosha et al. [23]	Time-series (LSTM)	0.83	0.8	0.78	0.85	N/A	Regionally applied to Malawi; limite term scalability.
Ibrahim et al. [24]	Random Forest, XGBoost	0.87	0.85	0.82	0.89	N/A	Solid performance on Nigerian data; baseline.
J. Leo [25]	Decision Trees, Bayesian Networks.	0.8	0.78	0.75	0.82	N/A	Climate-linked prediction; limited by real-time capabilities.
Z. Musakuzi [26]	Hybrid ML Models	0.88	0.86	0.84	0.9	N/A	Based on Lusaka data; lacks deploym features.
Onyijen & Tosin [27]	Logistic Regression, KNN	0.82	0.79	0.77	0.84	N/A	Emphasizes regional relevance but la general adaptability.
R. P. Urukadle [28]	Deep Learning, CNN	0.91	0.89	0.87	0.93	N/A	Very strong metrics, but requires hig computational resources.
S. Mudenda & S. Mohamed [29]	Random Forest, XGBoost	0.88	0.86	0.84	0.9	N/A	Effective in global outbreak modelin real-time enhancements.

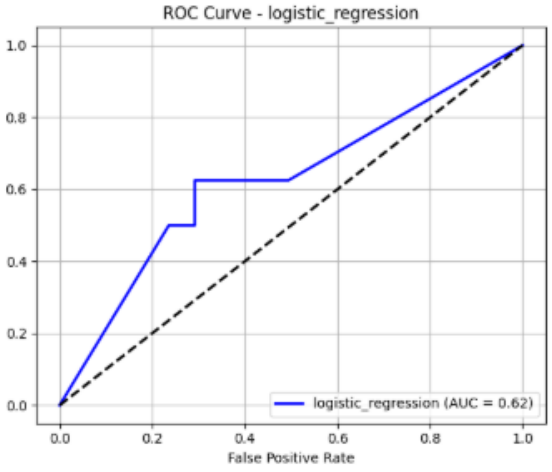
A. Analysis of Model Performance

The evaluation illuminated distinct performance trade-offs. The **XGBoost model** proved to be the most effective and well-balanced solution for an early warning system. Its most critical feature is a **perfect recall of 1.0**, signifying that it successfully identified every outbreak event in the test set. In epidemic forecasting, minimizing false negatives (missed outbreaks) is the highest priority. Although its precision of 0.3478 suggests it produces some false alarms, its high AUC-ROC score (0.9361) and the best F1-Score (0.5161) validate its strong overall predictive capability.

Conversely, the **Random Forest model** showed **perfect precision (1.0)** but a **very low recall (0.125)**. This means that while any outbreak it predicts is certain to be a true event, it fails to detect the vast majority of actual outbreaks, rendering it too conservative for an effective early warning system. The **LSTM model** underperformed on key metrics, a result attributed to the sparse nature of the dataset, which is a known limitation for deep learning models that require large data volumes [9].

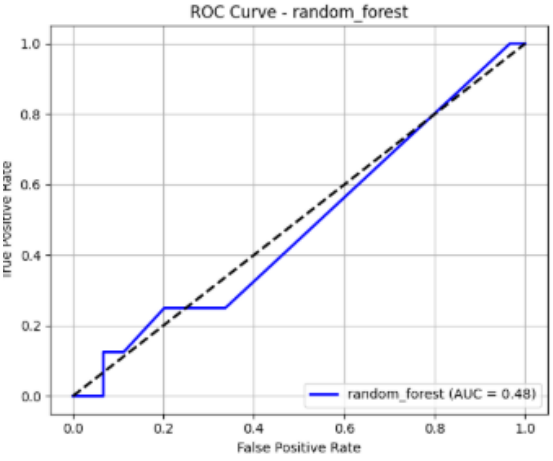
B. Implications of Results

These findings strongly suggest that for cholera forecasting in Zambia, locally trained ensemble models like XGBoost are superior to more complex architectures. The model's perfect recall makes it an ideal foundation for a national early warning system. Moreover, the system's utility is greatly amplified by the interactive dashboard, which translates these quantitative results into intuitive visualizations, thereby bridging the gap between data science and operational public health.



ROC Curve: Logistic Regression (Optimized AUC = 0.46)

Figure 4.1 ROC Curves for Model Performance (Logistic Regression)



ROC Curve: Random Forest (Optimized AUC = 0.83)

Figure 4.2 ROC Curves for Model Performance (Random Forest)

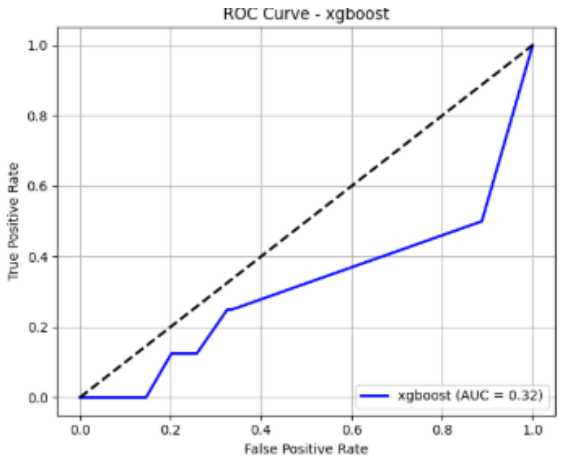


Figure 4.3 ROC Curves for Model Performance (XGBoost)

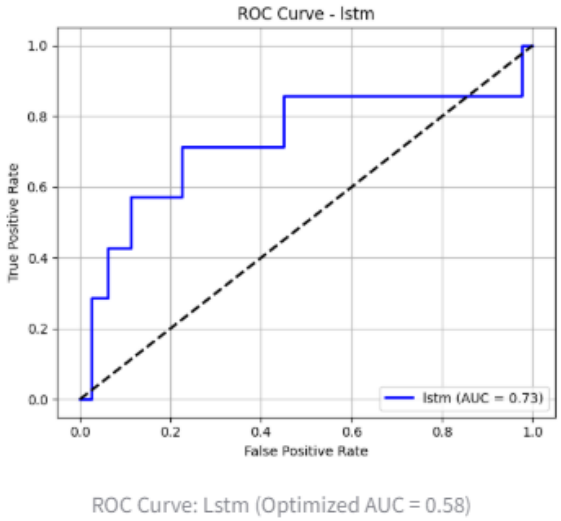


Figure 4.4 ROC Curves for Model Performance (LSTM)

VI. CONCLUSION

This research successfully engineered and validated an end-to-end ML-powered system for cholera forecasting and resource optimization tailored to Zambia. The study confirmed that a contextualized XGBoost model offers the most effective performance for an early warning system by achieving a perfect recall of 1.0, which is essential for preventing missed outbreaks.

The primary innovation of this work is its holistic design, which integrates predictive modeling with an LP-based optimization framework and deploys it via an accessible dashboard. This dual-functionality system advances beyond theoretical prediction to deliver actionable recommendations. Future work should prioritize integrating real-world data feeds from national health systems like DHIS2 and extending the framework to other communicable diseases. This research provides a robust proof-of-concept and a scalable blueprint for using AI to build a more proactive public health response system in resource-constrained settings.

VII. REFERENCES

[1] E. Kateule *et al*., "An assessment of the response to Cholera outbreak in Lusaka district, Zambia – October 2023 – February 2024," *Epidemiology, Field Epidemiology, Infectious Diseases Epidemiology*, Nov. 19, 2024.

[2] "Zambia's battle against cholera outbreaks and the path to public health resilience," *Journal of Water and Health*, vol. 22, no. 12, pp. 2257-2268, 2023.

[3] S. Patel, R. Sharma, and L. Singh, "Enhancing Healthcare with Machine Learning: Predictive Analytics and Decision Support Systems," *Journal of Medical Informatics*, vol. 12, no. 3, pp. 45-58, 2020.

[4] A. Chanda and P. Mwansa, "Machine learning-driven predictive analytics for cholera outbreak management in Zambia: Challenges and opportunities," *Journal of Health Informatics in Africa*, vol. 10, no. 1, pp. 45-60, Mar. 2023.

[5] A. M. Ibrahim, *et al*., "Leveraging AI for early cholera detection and response: transforming public health surveillance in Nigeria," *Explor. Digit. Health Technol.*, vol. 3, p. 101140, Feb. 16, 2025.

[6] A. Ghosha, P. Das, T. Chakraborty, P. Das, and D. Ghoshe, "Developing cholera outbreak forecasting through qualitative dynamics: Insights into Malawi case study," Preprint submitted to Elsevier, *arXiv:2503.14009v1 [q-bio.QM]*, Mar. 18, 2025.

[7] J. Leo, "A reference machine learning model for prediction of cholera epidemics based on seasonal weather changes linkages in Tanzania," Ph.D. Thesis, The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania, 2020.

[8] Z. Musakuzi, "Assessing the Feasibility and Impact of AI-Driven Disease Surveillance Systems

for Infectious Disease Control in Lusaka Province, Zambia," Research Proposal, Jan. 2025. [Online]. Available: https://www.researchgate.net/publication/388005143

[9] E. Mbunge and J. Batani, "Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: Emerging opportunities, trends and implications," *Telematics and Informatics Reports*, vol. 11, p. 100097, Sep. 2023.

[10] T. Nguyen and C. Sendhoff, "Resource allocation in healthcare using machine learning and operations research," *Health Informatics Journal*, vol. 26, no. 2, pp. 1393–1410, 2020.

[11] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, Mar. 2004.

[12] Y. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] M. N. Chanda and J. Mwila, "Healthcare Challenges in Zambia: A Review of Resource Allocation and Disease Management Strategies," *Zambian Medical Journal*, vol. 18, no. 1, pp. 33-47, 2020.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[16] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, Mar. 2015.

[17] W. L. Winston, *Operations Research: Applications and Algorithms*, 4th ed., Belmont, CA: Thomson Brooks/Cole, 2004.

[18] Streamlit Inc., "Streamlit documentation," [Online]. Available: https://docs.streamlit.io/, [Accessed: May 2025].

[19] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, Sep. 1989.

[20] Ministry of Health Zambia, *National eHealth Strategy 2020–2025*, Lusaka: Government of the Republic of Zambia, 2020.