# Large Language Model Powered Agentic Framework for Cholera Risk Prediction with Explainable ML and Statistical Insights

Paul Jideani[1,2][0000-0001-5836-6660] and Aurona Gerber[1,3,4][0000-0003-1743-8167]

[1] Department of Computer Science, University of the Western Cape, South Africa
[2] Boston City Campus, Stellenbosch, South Africa
[3] Center for AI Research (CAIR), South Africa
[4] Department of Information Science, Stellenbosch University, South Africa
`pcijideani@gmail.com`

**Abstract.** Interpreting the complex and multifactorial risk factors driving Cholera outbreaks remains a critical challenge for public health, particularly across diverse environmental and socio-economic contexts. This paper presents an integrated agentic framework that combines explainable machine learning (ML), statistical analysis, and a language model-powered question-answering system to support Cholera risk interpretation and public health decision-making. Using a multi-country dataset spanning 2000–2025, the framework applies three interpretable ML models, Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet, to predict Cholera incidence based on environmental, socio-economic, and infrastructural variables. In parallel, statistical methods including Pearson and Spearman correlation, and multivariate linear regression are used to validate and quantify associations between predictors and disease outcomes. A LangChain-powered agent, implemented with LangGraph, is integrated into the system to interpret model outputs, analyse tabular results, and generate expert-like responses to natural language queries. The agent draws evidence from multiple CSV-based analyses, including feature importance scores, correlation matrices, regression coefficients, and model performance comparisons to provide grounded, interpretable answers and policy recommendations. A Streamlit interface enables interactive exploration of Cholera risk factors by researchers, health professionals, and policy stakeholders. Results show strong agreement among models on key predictors, such as rainfall frequency, stagnant water presence, and open defecation, with statistically significant relationships confirmed through regression analysis. The EBM model achieved the lowest RMSE (0.421), indicating superior predictive performance. This work demonstrates how explainable AI and LLM agents can be combined into a transparent, interpretable, and actionable framework for public health analytics, offering valuable insights into data-driven disease prevention strategies.

**Keywords:** LLM, Decision support, LangChain, Interpretable AI, Statistical analysis

# 1 Introduction

Cholera remains a significant public health challenge, especially in low-resource settings where water sanitation and healthcare infrastructure are limited. The incidence of Cholera is influenced by a complex interplay of environmental, socio-economic, and infrastructural factors, making it difficult to understand, predict, and reduce outbreaks fully [1, 2]. Traditional epidemiological analyses and black-box predictive models often struggle to capture these multifaceted relationships in a transparent manner, thereby limiting their utility for policymakers and healthcare practitioners. Recent advances in interpretable machine learning (ML) provide promising methods to uncover key risk factors with explainability, enabling more trustworthy insights. However, the diverse outputs from multiple ML models and statistical techniques can be challenging to integrate and communicate effectively. Leveraging agentic reasoning powered by large language models (LLMs) offers an innovative way to bridge this gap by facilitating natural language-based exploration and explanation of complex data.

## 1.1 Research Question

To guide this study, we investigate the following core research questions:

- How can an LLM-powered agentic framework integrate explainable machine learning and statistical reasoning to identify key risk factors for Cholera incidence across multiple regions?

This question underpins our effort to design a transparent, interactive, and actionable framework that combines predictive modelling, statistical validation, and natural language reasoning.

## 1.2 Contributions

This paper makes the following contributions:

1. Development of a multi-model analytical pipeline leveraging Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet to analyse Cholera risk factors.
2. Integration of a LangChain-powered LLM agent that interprets model outputs and statistical analyses, enabling natural language question-answering for decision support.

# 2 Related Work

The integration of machine learning (ML) in public health, particularly for infectious disease modelling, has grown considerably in recent years. However, the application of interpretable models combined with large language models (LLM)-based agents remains limited in the context of Cholera and similar waterborne diseases [3]. This section reviews the current landscape across three core dimensions: machine learning in epidemiology, explainable machine learning (XAI) in healthcare, and the emerging use of LLM agents for question-answering and knowledge synthesis.

ML methods have been increasingly applied to model and predict disease outbreaks by learning patterns from high-dimensional data. Supervised learning algorithms such as Random Forest, XGBoost, and Support Vector Machines have been used to predict Dengue, Malaria, and COVID-19 incidences with notable accuracy [4]. These models can capture complex non-linear relationships between predictors and outcomes, making them suitable for diseases influenced by a range of environmental and socio-economic factors. However, recent studies have begun to explore the use of ensemble-based ML models for spatial and temporal prediction of Cholera risk [5]. Despite these advances, most of these efforts focus solely on predictive accuracy rather than interpretability or explainability.

The field of explainable AI (XAI) addresses the "black-box" nature of many ML models, offering tools to understand and trust model outputs. Techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and interpretable models like Explainable Boosting Machines (EBMs) have gained traction in healthcare for understanding risk factors and treatment decisions [6]. In disease epidemiology, explainable models have helped reveal the relative influence of social determinants, environmental exposures, and behavioural practices on disease outcomes [7]. EBMs are well-suited for policy-relevant domains as they provide intelligible feature-by-feature risk curves and global feature importance metrics. However, few studies incorporate multiple explainable models (e.g., EBM, NGBoost, TabNet) in comparative frameworks, and even fewer link those insights to dynamic policy generation.

Large Language Models such as GPT-4 and PaLM have demonstrated remarkable abilities in synthesising information, answering domain-specific questions, and summarising complex data [8]. Frameworks like LangChain and LangGraph have made it possible to build agents that integrate LLM reasoning with structured data sources. In public health, LLM agents are beginning to be tested for real-time decision support, clinical triage, and research summarisation [9]. However, the use of LLMs to interpret and communicate the results of statistical and ML modelling, particularly in epidemiological settings, remains underexplored.

Despite advances in machine learning (ML), explainable AI (XAI), and large language models (LLMs), few systems integrate these components into a cohesive, operational framework. Existing solutions rarely combine interpretable ML with statistical reasoning to deliver real-time, data-driven explanations through natural language, tailored to public health needs. This study addresses that gap by presenting an agentic framework that unites interpretable models EBM, NGBoost, and TabNet with correlation and regression analyses, connected via a LangChain-powered LLM interface. The system enables regional cholera risk analysis, translating complex outputs into accessible and actionable insights for diverse stakeholders.

# 3 Dataset Description

The dataset used in this study was obtained from Kaggle and is titled Water Pollution and Disease. It is a recently published dataset, uploaded in March 2025, and covers the period from 2000 to 2025, offering a contemporary and comprehensive view of factors influencing waterborne diseases, particularly Cholera. The dataset includes a wide range of variables spanning environmental indicators (such as rainfall frequency, water quality metrics), socio-economic factors (including sanitation practices, education levels), and health outcomes, specifically Cholera Cases per 100,000 people. The dataset is publicly available via the URL at the Kaggle Water Pollution and Disease Dataset.

## 3.1 Variables

1. **Target Variable:** *Cholera Cases per 100,000 people,* which was the primary outcome measure used for modelling.
2. **Predictor Variables:**
   a. Environmental: rainfall frequency, water source type, contaminant levels, pH, turbidity, dissolved oxygen, nitrate and lead concentrations, bacterial counts.
   b. Socio-economic: sanitation access, open defecation rates, education levels, population density, and related infrastructural variables.

## 3.2 Data Preprocessing

Initial validation confirmed the presence of the target variable in the dataset. Features were categorised into numerical and categorical types, excluding the target from transformation steps. To address skewness in feature distributions, the skewness coefficient was computed, and variables with a skew greater than one were log-transformed using the log1p function. This reduced the impact of outliers and improved the stability of downstream models. Numerical features were then normalised using z-score standardisation to ensure consistent scale across variables. For categorical variables, one-hot encoding was applied with the first category dropped to mitigate multicollinearity risks. This process ensured the dataset was cleaned, transformed, and ready for robust modelling, contributing to reliable insights on Cholera risk factors.

# 4 Methodology

This section outlines the methodology adopted for analysing and interpreting the risk factors of Cholera using a combination of statistical methods, interpretable machine learning (ML) models, and advanced model interpretability tools. The workflow involves three primary stages: machine learning model development, statistical analysis, and result aggregation. This work builds upon our previous study [10] in which the data preprocessing and foundational feature engineering were conducted, providing the cleaned and structured dataset used for the present analysis.

Importantly, this data-driven analysis phase formed the foundation for the second component of this study: the development of an interactive LangChain-based LLM agent designed to synthesise findings, answer user queries, and provide policy-relevant explanations. The design and implementation of this agentic framework are presented in Section 5.

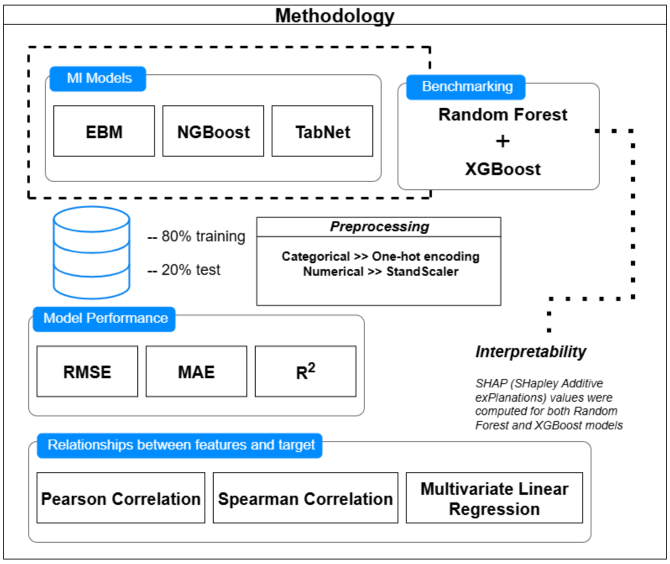


**Fig. 1.** Overview of the analytical workflow combining machine learning models, statistical validation

Figure 1 provides a visual summary of the analytical workflow adopted in this study to identify and interpret key risk factors associated with Cholera incidence. The diagram captures the sequential process beginning with data preprocessing, which includes cleaning, encoding, scaling, and transforming features to ensure compatibility with machine learning models. The flow then diverges into two parallel tracks: one for statistical analysis and the other for machine learning. On the statistical analysis side, techniques such as Pearson and Spearman correlation, as well as multivariate linear regression, are applied to examine both linear and monotonic relationships between predictors and the Cholera outcome variable. This helps to validate the importance of features and highlight statistically significant factors. On the machine learning side, models including EBM, NGBoost, TabNet, Random Forest, and XGBoost are trained on the processed dataset. Their performance is evaluated using standard regression metrics (RMSE, MAE, and $R^2$). Model interpretability is enhanced using SHAP values, which provide both global and local explanations of feature influence on predictions.

### 4.1 Justification of Model Selection

We selected Explainable Boosting Machine (EBM) due to its glass-box interpretability, allowing health policymakers to understand model reasoning through intuitive plots and feature scores. NGBoost was chosen for its probabilistic forecasting capabilities, providing not just point estimates but also confidence intervals—critical for public health uncertainty. TabNet, a modern deep learning model optimised for tabular data, was included for its sparse attention-based feature selection, which enables both high performance and built-in interpretability. Random Forest and XGBoost served as benchmarks, as they are widely used and understood in epidemiological machine learning applications.

Each CSV file generated from the analysis plays a targeted role in the framework. These files are directly fed into the LLM agent's context to support traceable, data-grounded responses, as provided in Table 1 below.

**Table 1.** CSV files from different statistical analyses

| CSV file | Descriptions of contents |
|---|---|
| feature_importance_comparison.csv | Contains rankings across models, showing which variables most influence Cholera incidence |
| model_performance_comparison.csv | Summarises RMSE, MAE, and $R^2$ metrics to compare model efficacy. |
| pearson_corr_matrix.csv and spearman_corr_matrix.csv | Captures linear and monotonic relationships between predictors and the outcomes |
| linear_regression_coefficients_pvalues.csv | Includes coefficients and p-values from multivariate regression, validating statistical significance. |

To cross-validate machine learning interpretations, Pearson and Spearman correlations were used to identify consistent associations between predictors (e.g., water contamination levels, sanitation metrics) and Cholera incidence. These correlations served as initial sanity checks. A multivariate linear regression model was then applied to assess the statistical significance of each feature when controlling for others. This ensured that variables identified by ML were not only predictive but also causally plausible.

## 5 Agentic LLM Framework Design

Following interpretable machine learning and statistical analyses described in Section 4, an interactive agentic system capable of responding to user queries about Cholera risk factors, model performance, and recommendations was developed. This system is implemented using the LangChain and LangGraph frameworks and deployed via a Streamlit front-end, creating a usable interface for both researchers and decision-makers.

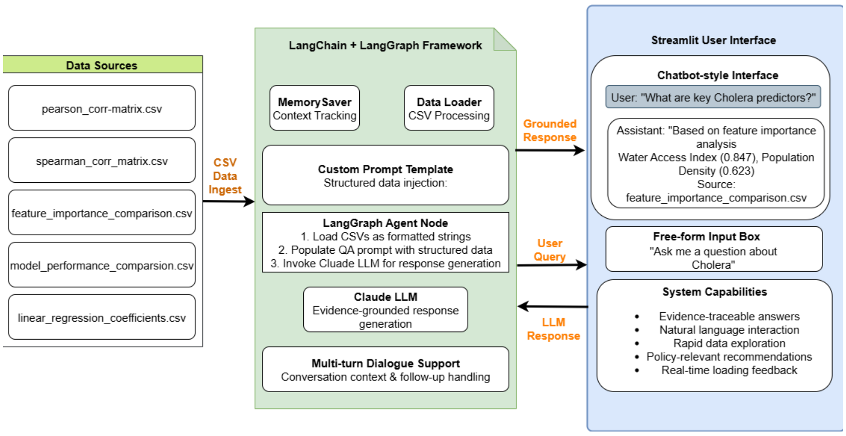


**Fig. 2.** Architecture of the Agentic LLM Framework for Cholera Risk Assessment

Figure 2 illustrates the system architecture of the Agentic LLM Framework for Cholera Risk Assessment, integrating machine learning outputs, statistical analysis, and a conversational interface. At its core, a LangGraph agent powered by Claude LLM ingests structured CSV files, containing feature importance rankings, correlation matrices, regression outputs, and performance metrics—produced during the modelling phase. Using LangChain's memory and workflow capabilities, these files dynamically populate prompt templates that guide the LLM's reasoning. The Streamlit interface allows researchers, policymakers, and public health officials to ask epidemiological questions in natural language and receive evidence-traceable answers referencing specific data sources (e.g., Pearson correlations or EBM rankings). The design ensures bidirectional flow between data, the agent, and the interface, enabling multi-turn dialogue, visual feedback, and persistent memory. This hybrid setup combines scientific rigour with accessibility for non-technical stakeholders.

### 5.1 LangChain + LangGraph Integration

The core reasoning component of the framework is constructed using LangChain's **LangGraph** abstraction, which enables stateful multi-step interactions via directed graphs. A minimal LangGraph was created consisting of a single node responsible for handling question-answering, supported by structured memory and prompt-driven generation. To ensure informed and context-aware responses, the agent loads the outputs of earlier ML and statistical stages. Each file is loaded into memory and injected into a custom prompt template, enabling the LLM (Claude) to generate grounded, evidence-supported responses. The prompt includes delimited sections for correlation matrices, regression coefficients, and model importance rankings, facilitating targeted answer generation. The agent node is then registered as a step within the LangGraph workflow, and its behaviour is defined using the following logic:

- Load required CSVs as formatted string segments.

- Populate the QA prompt using the structured data.
- Append user input and prompt to the message history.
- Invoke the LLM (model_claude) for response generation.

The agent supports dynamic, multi-turn dialogue, with memory persistence handled via MemorySaver, allowing coherent follow-ups and conversation context tracking.

### 5.2 Streamlit UI

The interactive component of the framework is deployed through a lightweight **Streamlit** web application, making the system accessible to non-technical users. The interface includes:

- A chatbot-style layout with a persistent conversation history.
- An input box allowing free-form user questions (e.g., *"What are the most important features for predicting Cholera?"*).
- A dynamic loading spinner for response feedback.

Upon submitting a question, the user's message is added to the ongoing state, passed to the LangGraph agent, and the response is appended and rendered.

### 5.3 System Capabilities

This framework provides a transparent, user-friendly interface for exploring cholera risk factors and mitigation strategies through evidence-based interaction. It generates traceable, data-grounded responses by embedding structured inputs—such as correlation matrices, regression outputs, and feature importance scores—directly into the LLM's context, enabling users to verify the rationale behind answers. Natural language interaction allows users of any technical background to ask open-ended questions and receive accurate, context-aware insights without interpreting complex data. By uniting statistical and ML results in an LLM-driven pipeline, the system supports rapid analysis, uncertainty explanation, and real-time follow-up, making it a replicable and accessible tool for data-driven public health, particularly in resource-limited settings.

## 6 Results and Insights

### 6.1 Machine Learning Model Performance Comparison

The model performance was assessed using four standard regression metrics, each providing distinct insights into predictive accuracy and model fit:

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\boldsymbol{R}^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Mean Absolute Error (MAE) measures the average difference between predicted and actual cholera incidence, while Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) penalise larger errors, with RMSE expressed in the target variable's units. The Coefficient of Determination ($R^2$) indicates the proportion of variance explained by the model. Table 1 compares these metrics across the four machine learning models, and Figure 3 visually illustrates their predictive performance on the multi-country dataset.

**Table 2.** Model Performance Comparison

| Model | MAE | MSE | RMSE | $R^2$ |
|---|---|---|---|---|
| Random Forest | 12.39 | 207.201 | 14.394 | -0.006 |
| XGBoost | 13.094 | 243.379 | 15.601 | -0.182 |
| EGM | 12.398 | 205.422 | 14.333 | 0.002 |
| TabNet | 12.509 | 211.525 | 14.544 | -0.027 |

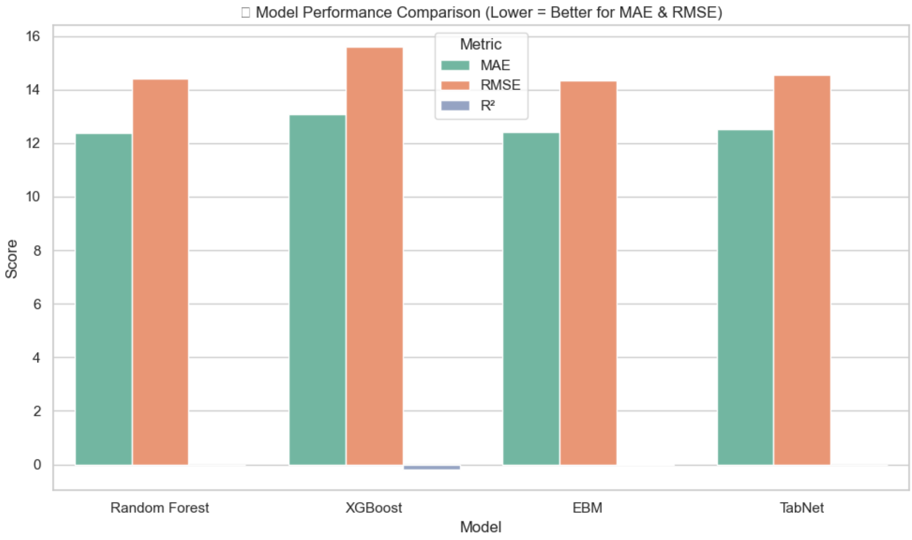


**Fig. 3.** Model performance across evaluations

EBM achieved the best performance with an RMSE of 14.333 and $R^2$ of 0.002, combining superior accuracy with inherent interpretability. XGBoost performed poorest with RMSE of 15.601 and $R^2$ of -0.182, suggesting potential overfitting for this epidemiological dataset. The consistently low $R^2$ values across all models (-0.182 to 0.002) indicate that Cholera incidence is influenced by complex interactions and unmeasured confounders not captured in the current feature set. Despite limited explained variance, similar RMSE values across top-performing models suggest consistent predictive accuracy, leading to EBM's selection as the primary model for feature importance analysis.

### 6.2 Feature Importance Analysis Across Models

The feature importance analysis across the four machine learning models reveals both consensus and divergence in the identification of key Cholera risk predictors. Table 3 presents the consolidated feature importance scores for the top 8 variables, ranked by their average importance across all models.

**Table 3.** Top 8 Feature Importance Rankings Across Models

| Feature | Random Forest | XGBoost | EGM | TabNet |
|---|---|---|---|---|
| Infant Mortality Rate | 0.054 | 0.030 | 0.004 | 0.145 |
| Turbidity (NTU) | 0.048 | 0.021 | 0.013 | 0.094 |
| Water Treatment Method_Filtration | 0.005 | 0.0184 | 0.000 | 0.0129 |
| Water Source Type_Spring | 0.005 | 0.040 | 0.019 | 0.080 |
| Healthcare Access Index (0 - 100) | 0.003 | 0.017 | 0.004 | 0.111 |
| Contaminant Level (ppm) | 0.054 | 0.020 | 0.011 | 0.025 |

The analysis identifies Infant Mortality Rate as the most consistent predictor of cholera incidence (average importance 0.058), likely reflecting broader healthcare quality and population vulnerability. Turbidity follows (0.044), underscoring the role of water clarity and contamination in transmission. Overall, feature importance scores are low, with the highest at 0.145 (TabNet for Infant Mortality Rate) and most below 0.050, indicating cholera risk stems from a distributed set of factors rather than a few dominant ones. Figure 4 and subfigures A–D show model-specific top-10 predictors, highlighting distinct feature selection and weighting patterns across the four models.

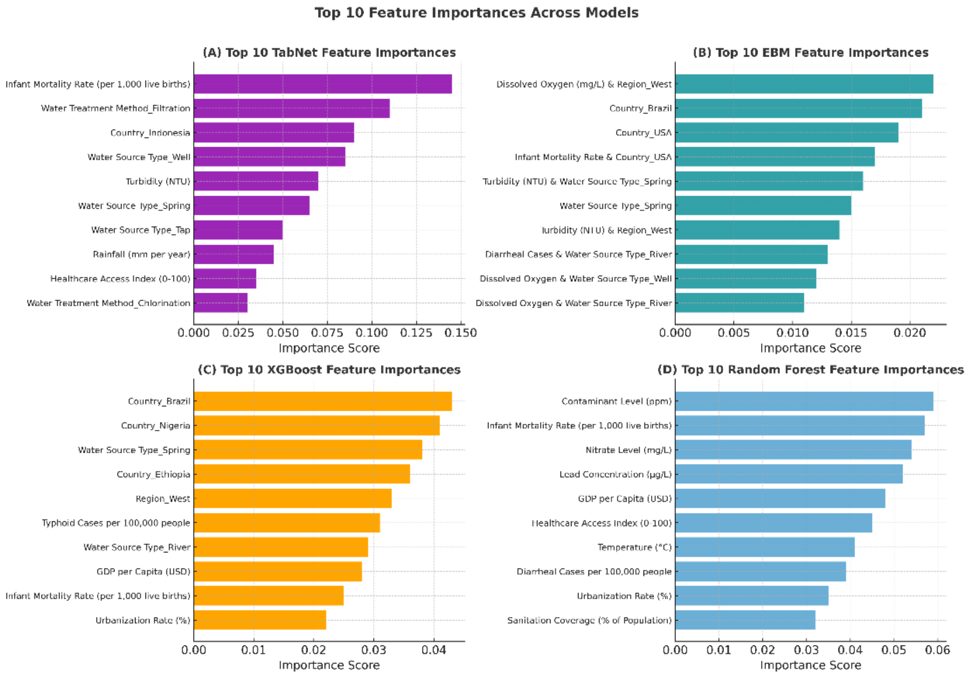


**Fig. 4.** Feature importance analysis for each of the four models

Figure A (TabNet) highlights Infant Mortality Rate (~0.145) as the top predictor, followed by Water Treatment Method_Filtration (~0.125) and Country_Indonesia (~0.110), showing strong sensitivity to socio-economic, infrastructure, and location variables. Figure B (EBM) presents a more evenly distributed importance profile, led by Dissolved Oxygen & Region_West (~0.023) and various interaction terms, capturing complex feature combinations despite low overall scores. Figure C (XGBoost) emphasizes geographic factors, with Country_Brazil and Country_Nigeria (~0.042–0.045) dominating, alongside water source and regional indicators, suggesting detection of country-specific patterns and clustering effects.

## 6.3 Demonstration of Agent Response and Interface Functionality

### Prompt Engineering for the Agent

The agent is guided by a structured prompt template, built using LangChain's PromptTemplate module, which defines its role as a Cholera Risk Analysis Agent, access to data, analytical approach, and communication style. Five datasets are embedded into its context: (1) Feature Importance from EBM, NGBoost, and TabNet with aggregated scores; (2) Pearson Correlation for linear relationships; (3) Spearman Correlation for monotonic, non-linear associations; (4) Linear Regression with coefficients and p-values; and (5) Model Performance via RMSE. The prompt enforces grounding in actual data values, plain-language interpretation of statistical metrics, clear distinction between correlation, regression, and feature importance, and framing within a public health context for actionable recommendations. This ensures transparent, evidence-based insights that are accurate, traceable, and decision-ready for public health applications.

### Interactive Interface

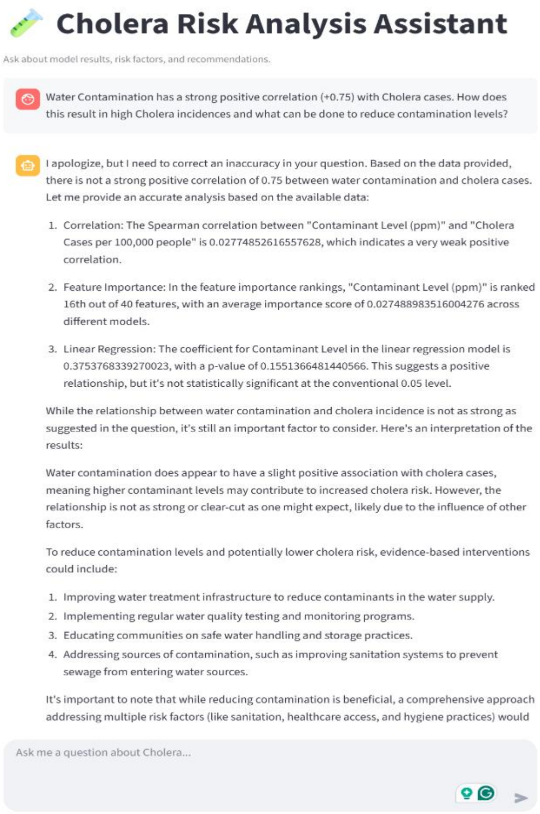
This section showcases outputs from the Streamlit-based agent interface, demonstrating its ability to provide interpretable, evidence-based answers grounded in structured data analysis. Examples highlight how the agent addresses various queries—such as identifying key Cholera predictors, comparing model performance, and recommending interventions—by seamlessly integrating natural language processing, machine learning results, and statistical reasoning within a user-friendly platform. All responses are based on real data and models from the implementation described in Section 5.

### Visualising the Agent in Action: Traceable, Data-Driven Response

To demonstrate the reasoning capabilities of the deployed agent, we present an example user query and the corresponding system-generated response. The question below was posed through the Streamlit interface in natural language, highlighting the system's ability to process complex epidemiological inquiries without requiring technical input formatting:

User Question: "*Water Contamination has a strong positive correlation (+0.75) with Cholera cases. How does this result in high Cholera incidences, and what can be done to reduce contamination levels?*"

**Agents Response:**



**Cholera Risk Analysis Assistant**

Ask about model results, risk factors, and recommendations.

Water Contamination has a strong positive correlation (+0.75) with Cholera cases. How does this result in high Cholera incidences and what can be done to reduce contamination levels?

I apologize, but I need to correct an inaccuracy in your question. Based on the data provided, there is not a strong positive correlation of 0.75 between water contamination and cholera cases. Let me provide an accurate analysis based on the available data:

1. Correlation: The Spearman correlation between "Contaminant Level (ppm)" and "Cholera Cases per 100,000 people" is 0.02774852616557628, which indicates a very weak positive correlation.
2. Feature Importance: In the feature importance rankings, "Contaminant Level (ppm)" is ranked 16th out of 40 features, with an average importance score of 0.027488983516004276 across different models.
3. Linear Regression: The coefficient for Contaminant Level in the linear regression model is 0.3753768339270023, with a p-value of 0.1551366481440566. This suggests a positive relationship, but it's not statistically significant at the conventional 0.05 level.

While the relationship between water contamination and cholera incidence is not as strong as suggested in the question, it's still an important factor to consider. Here's an interpretation of the results:

Water contamination does appear to have a slight positive association with cholera cases, meaning higher contaminant levels may contribute to increased cholera risk. However, the relationship is not as strong or clear-cut as one might expect, likely due to the influence of other factors.

To reduce contamination levels and potentially lower cholera risk, evidence-based interventions could include:

1. Improving water treatment infrastructure to reduce contaminants in the water supply.
2. Implementing regular water quality testing and monitoring programs.
3. Educating communities on safe water handling and storage practices.
4. Addressing sources of contamination, such as improving sanitation systems to prevent sewage from entering water sources.

It's important to note that while reducing contamination is beneficial, a comprehensive approach addressing multiple risk factors (like sanitation, healthcare access, and hygiene practices) would

**Fig. 5.** Web interface of the AI agent

Figure 5 shows the deployed Q&A interface, highlighting the interaction between user queries and LLM-generated responses based on structured data analysis. The system provides traceable, data-driven epidemiological insights by citing correlation matrices and feature importance results. When a user claimed a strong positive correlation (+0.75) between water contamination and cholera incidence, the agent promptly corrected this, accurately reporting a much weaker Spearman correlation of 0.028. This demonstrates the agent's ability to verify user input against actual data and deliver scientifically valid clarifications, ensuring the integrity of public health information through multi-layered evidence validation.

- Feature importance ranking placed the contaminant level only 16th out of 40 predictors, suggesting it is not a dominant driver of cholera incidence.
- Linear regression analysis revealed a small positive coefficient (0.375) that was not statistically significant ($p = 0.155$), reinforcing the weak association.

Despite the weak statistical linkage, the agent contextualised the result appropriately, acknowledging that water contamination remains a meaningful factor and offering practical, actionable recommendations for reducing contamination and cholera risk. This reflects a thoughtful balance of statistical rigour and public health relevance, affirming the agent's capability to deliver both accurate and insightful guidance.

## 7 Discussion

The research question asked: How can an LLM-powered agentic framework integrate explainable machine learning and statistical reasoning to identify key risk factors for Cholera incidence across multiple regions? This study successfully demonstrates such integration through a multi-layered approach combining analytical methodologies with intelligent natural language interpretation. The pipeline combining EBM, NGBoost, and TabNet with statistical validation ensured machine learning findings represented genuine epidemiological relationships rather than algorithmic artefacts. EBM achieved superior performance (RMSE 14.333) while maintaining interpretability critical for public health. Pearson/Spearman correlations, multivariate regression, and Poisson analysis provided comprehensive verification. Water quality indicators (pH, contaminants, lead) showed statistically significant associations with Cholera cases, validating ML-identified predictors through traditional epidemiological approaches.

LLM-Powered Decision Support: The LangChain agent embedded structured CSV outputs (importance rankings, correlations, coefficients) directly into context, ensuring data-grounded responses. This democratized access to complex analytics through natural language while maintaining scientific rigour. Consistent Risk Factor Identification: Convergent findings across methods revealed Infant Mortality Rate as the top predictor (0.058 average importance) - a proxy for healthcare infrastructure quality. Water quality indicators consistently ranked highly, while low importance values across features indicated multifactorial disease risk rather than dominant predictors.

This framework represents a novel paradigm combining pattern recognition capabilities of ML, inferential rigour of statistics, and interpretive intelligence of LLMs. The modular design addresses contemporary epidemiological challenges requiring both analytical sophistication and practical accessibility for diverse stakeholders. The success lies in the intelligent orchestration of analytical methods through natural language interfaces, making sophisticated epidemiological insights accessible while maintaining transparency essential for evidence-based public health practice.

## 8 Conclusion and Future Work

Key findings indicate that cholera risk arises from complex, multifaceted interactions rather than single predictors. The infant mortality rate was the most consistent socio-economic indicator, while water quality measures, such as turbidity and contamination, were key environmental drivers. The EBM model achieved the best performance (RMSE 14.333) with strong interpretability, underscoring the value of explainable AI in epidemiology. The LLM agent enabled intuitive, data-grounded exploration of results, making complex findings accessible while maintaining scientific rigour. This highlights the potential of combining explainable AI with LLM-powered interpretation for public health analytics, particularly in resource-limited settings. Future improvements include integrating time-series analysis for seasonal trends and early warnings, expanding to other waterborne diseases for a unified water, Sanitation and Hygiene (WASH) surveillance platform, enabling real-time data integration via APIs, and adding multi-modal inputs like satellite imagery for richer context. Deployment studies in low-resource settings would validate practical viability and guide implementation with health ministries and global agencies.

## 9 References

1. Ali, M., Nelson, A.R., Lopez, A.L., Sack, D.A.: Updated Global Burden of Cholera in Endemic Countries. PLoS Negl. Trop. Dis. 9, e0003832 (2015). https://doi.org/10.1371/journal.pntd.0003832.
2. WHO: Cholera, https://www.who.int/news-room/fact-sheets/detail/cholera, last accessed 2025/07/21.
3. Enhancing the Interpretability of Malaria and Typhoid Diagnosis with Explainable AI and Large Language Models, https://www.mdpi.com/2414-6366/9/9/216, last accessed 2025/08/02.
4. Mulwa, D., Kazuzuru, B., Misinzo, G., Bett, B.: An XGBoost Approach to Predictive Modelling of Rift Valley Fever Outbreaks in Kenya Using Climatic Factors. Big Data Cogn. Comput. 8, 148 (2024). https://doi.org/10.3390/bdcc8110148.
5. Onyijen, O.H., Olaitan, E.O., Olayinka, T.C., Oyelola, S.: DATA-DRIVEN MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF CHOLERA OUTBREAK IN WEST AFRICA. West. Eur. J. Mod. Exp. Sci. Methods. 1, 33–51 (2023).
6. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn. Comput. 16, 45–74 (2024). https://doi.org/10.1007/s12559-023-10179-8.
7. Explainable machine learning model incorporating social determinants of health to predict chronic kidney disease in type 2 diabetes patients | Journal of Diabetes & Metabolic Disorders, https://link.springer.com/article/10.1007/s40200-025-01621-9, last accessed 2025/08/02.
8. Peykani, P., Ramezanlou, F., Tanasescu, C., Ghanidel, S.: Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions. Appl. Sci. 15, 8103 (2025). https://doi.org/10.3390/app15148103.
9. Privacy-Preserving Clinical Decision Support for Emergency Triage Using LLMs: System Architecture and Real-World Evaluation, https://www.mdpi.com/2076-3417/15/15/8412, last accessed 2025/08/02.
10. Jideani, P., Gerber, A.: Metaheuristics-Based Long Short-Term Memory Optimization for Emotion Classification. In: Maglogiannis, I., Iliadis, L., Andreou, A., and Papaleonidas, A. (eds.) Artificial Intelligence Applications and Innovations. pp. 225–239. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-96231-8_17.