# An Agentic LLM-Powered Framework for Cholera Risk Prediction and Actionable Insights from Explainable ML and Statistical Insights

# **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, analyze 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 for data-driven disease prevention strategies.

# **1.0 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 fully understand, predict, and mitigate outbreaks [1, 2]. Traditional epidemiological analyses and black-box predictive models often struggle to capture these multifaceted relationships transparently, limiting their utility for policymakers and health practitioners [3].

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 difficult 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 [4].

## **1.1 Research Questions**

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:

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

# **2.0 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 [5]. 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.

## **2.1 Machine Learning for Epidemiological Modelling**

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 [6]. 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. In Cholera research specifically, statistical models such as logistic regression and time-series models have historically been used to forecast outbreaks [7]. However, recent studies have begun to explore the use of ensemble-based ML models for spatial and temporal prediction of Cholera risk [8]. Despite these advances, most of these efforts focus solely on predictive accuracy rather than interpretability or explainability.

## **2.2 Explainable Machine Learning in Health Contexts**

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 [9]. In disease epidemiology, explainable models have helped reveal the relative influence of social determinants, environmental exposures, and behavioural practices on disease outcomes [10]. EBMs, in particular, 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.

## **2.3 LLM Agents for Knowledge Synthesis and QA**

Large Language Models such as GPT-4 and PaLM have demonstrated remarkable abilities in synthesizing information, answering domain-specific questions, and summarizing complex data [11]. 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 summarization [12]. However, the use of LLMs to interpret and communicate the results of statistical and ML modelling- particularly in epidemiological settings remains underexplored.

While machine learning (ML) and explainable artificial intelligence (XAI) methods have seen significant advancements, and large language models (LLMs) offer promising capabilities for interaction and synthesis, there remains a notable gap in the development of integrated systems that bring these components together in a cohesive, operational framework. Specifically, few existing solutions effectively combine interpretable ML modelling with statistical reasoning, deliver real-time, data-driven explanations through a natural language interface, and translate complex analytical outputs into localized insights relevant for public health interventions. This study addresses that gap by proposing an agentic framework that unifies explainable ML models—including Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet with statistical correlation and regression analyses. These components are further connected through a LangChain-powered LLM interface, which interprets model outputs and statistical findings in conversational form. The resulting system enables Cholera risk analysis across regions while enhancing accessibility to analytical insights for diverse stakeholders.

# **3.0 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 at [Kaggle Water Pollution and Disease Dataset](https://www.kaggle.com/datasets/khushikyad001/water-pollution-and-disease/data).

## **3.1 Variables**

* **Target Variable:** *Cholera Cases per 100,000 people* — the primary outcome measure used for modelling.
* **Predictor Variables:** 
  + Environmental: rainfall frequency, water source type, contaminant levels, pH, turbidity, dissolved oxygen, nitrate and lead concentrations, bacterial counts.
  + 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 categorized 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 1 were log-transformed using the log1p function. This reduced the impact of outliers and improved the stability of downstream models. Numerical features were then normalized using **z-score standardization** 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.

# **4.0 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.

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.

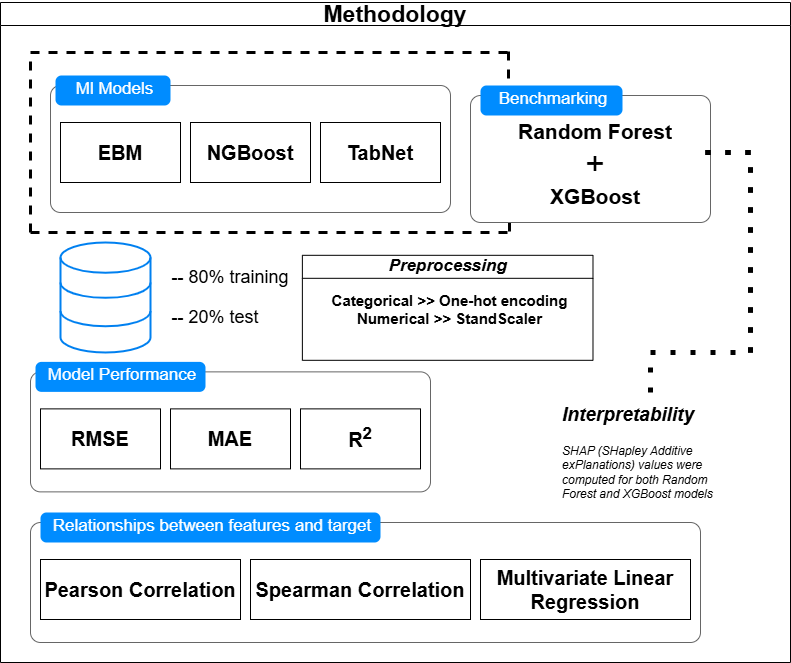


Figure 1 Overview of the analytical workflow combining machine learning models, statistical validation

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. These help 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 R²). 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 optimized 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.

* feature\_importance\_comparison.csv: Contains rankings across models, showing which variables most influence Cholera incidence.
* model\_performance\_comparison.csv: Summarizes RMSE, MAE, and R² metrics to compare model efficacy.
* pearson\_corr\_matrix.csv and spearman\_corr\_matrix.csv: Capture linear and monotonic relationships between predictors and the outcome.
* 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.

# **5.0 Agentic LLM Framework Design**

Following the interpretable machine learning and statistical analyses described in Section 4, we developed an interactive agentic system capable of responding to user queries about Cholera risk factors, model performance, and recommendations. 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.

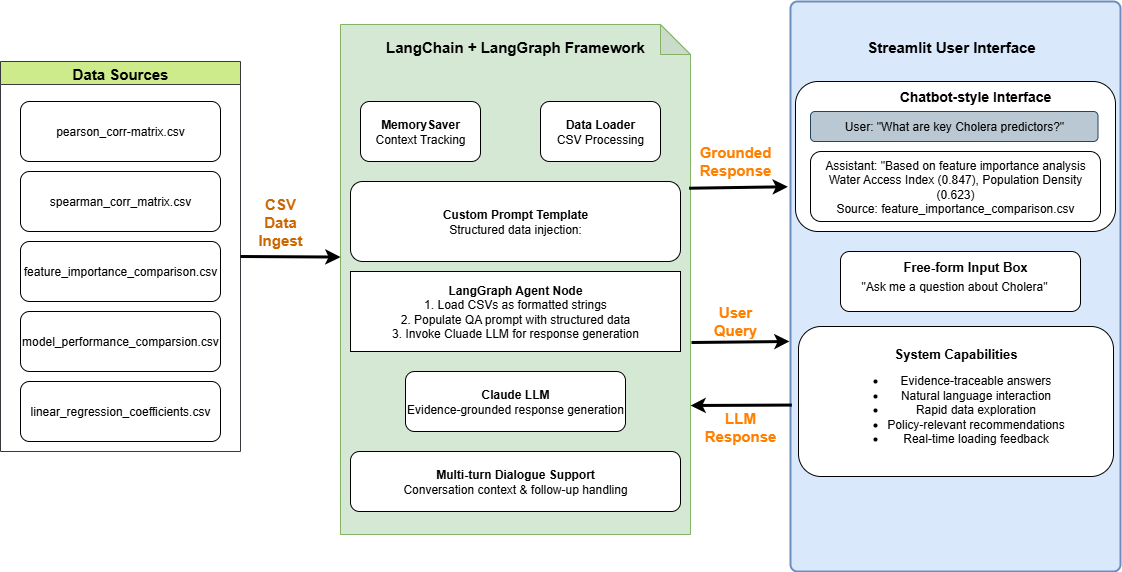


Figure 2 Architecture of the Agentic LLM Framework for Cholera Risk Assessment

Figure 2 shows the system architecture of the Agentic LLM Framework for Cholera Risk Assessment, illustrating the end-to-end integration of machine learning outputs, statistical analysis, and a user-facing conversational interface. At the core of the architecture is a LangGraph agent powered by Claude LLM, designed to ingest structured CSV files containing feature importance rankings, correlation matrices, regression outputs, and model performance metrics. These files, generated from the earlier modelling phase, are accessed through LangChain's memory and utility functions to dynamically populate a prompt template that guides the language model's reasoning. The system leverages LangChain’s structured state handling and LangGraph’s workflow capabilities to manage the flow of user input, agent response, and contextual memory across dialogue turns.

The Streamlit interface provides a clean, interactive platform where users—whether researchers, policymakers, or public health officials—can pose epidemiological and policy-relevant questions using natural language. The framework supports evidence-traceable answers, where responses reference specific CSV-backed insights, such as "according to the Pearson correlation matrix…" or "as seen in the feature importance rankings from EBM…". The interface also includes visual feedback like loading indicators, conversation history, and persistent memory to enable multi-turn interactions. The diagram further depicts how the components are interconnected: CSV data feeds into the agent, which in turn communicates with the user interface bidirectionally. This hybrid design enhances accessibility for non-technical stakeholders while ensuring scientific robustness and reproducibility of insights.Top of Form

Bottom of Form

## **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 clearly 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 integrated framework enables a robust and user-friendly interface for evidence-grounded exploration of Cholera risk factors and mitigation strategies. One of the core strengths of the system lies in its ability to produce evidence-traceable answers. Each response generated by the agent is not merely speculative or generic; rather, it is grounded in structured data inputs—including correlation matrices, regression outputs, and machine learning-derived feature importance scores—formatted as CSVs. These files are loaded and programmatically embedded within the prompt context supplied to the LLM, ensuring that the output references actual model results and statistical relationships. This traceability ensures that users can verify and interpret the rationale behind the answers, lending credibility and transparency to the system.

In addition, the framework supports natural language interaction, allowing users to pose questions in an intuitive, open-ended format. Whether a policymaker seeks insights into the top environmental contributors to Cholera outbreaks, or a health researcher wants to compare model performances, the system parses the query and retrieves relevant content without requiring the user to understand the technical structure of the underlying data. This makes the agent accessible to non-specialist stakeholders and promotes interdisciplinary usability.

Furthermore, the system supports rapidexploration of complex analytical outputs. By synthesising diverse statistical and machine learning results into a single LLM-driven pipeline, the agent can quickly summarise key predictors, explain uncertainty, and offer evidence-based recommendations. Users can ask follow-up questions in the same session, enabling dynamic and multi-faceted inquiry. For instance, after identifying that poor sanitation and high nitrate levels are key predictors, a user could ask how these variables vary across countries or how mitigation might reduce predicted Cholera incidence. The system responds fluidly, supported by memory persistence and structured prompting logic.

Overall, this hybrid design exemplifies a novel application of agentic LLMs in the epidemiological domain. By combining explainable machine learning, statistical validation, and conversational AI, the framework offers a powerful tool for data-driven public health insights. It provides a replicable, transparent, and interactive platform for guiding policy decisions, especially in resource-limited settings where timely and interpretable guidance is critical.

# **6.0 Results and Insights**

## **6.1 Machine Learning Model Performance Comparison**

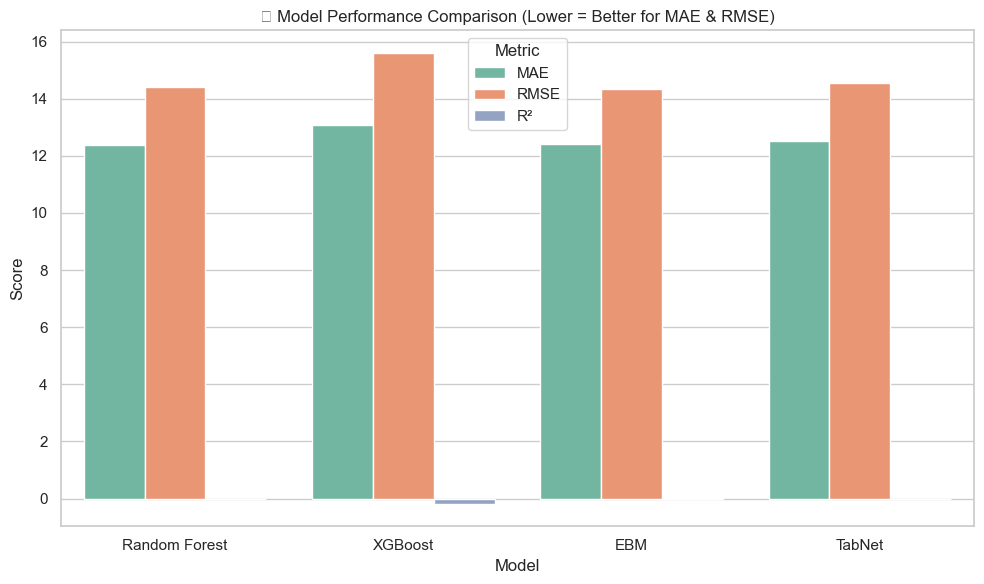
The model performance was assessed using four standard regression metrics, each providing distinct insights into predictive accuracy and model fit:

Mean Absolute Error (MAE): measures the average magnitude of errors between predicted and actual values. Where ​ ​ represents the actual Cholera incidence, ​ is the predicted value, n is the number of observations. Mean Squared Error (MSE) calculates the average of squared differences between predicted and actual values, penalizing larger errors more heavily. Root Mean Squared Error (RMSE) represents the square root of MSE, providing error measurement in the same units as the target variable. Coefficient of Determination (R²) quantifies the proportion of variance in the dependent variable explained by the model. Where SSres is the sum of squared residuals, SStot is the total sum of squares, and ​ is the mean of observed values.

The performance evaluation of the four machine learning models reveals nuanced differences in their ability to predict Cholera incidence across the multi-country dataset. Table 1 presents the comparative performance metrics, while Figure 6 provides a visual comparison of model performance across key evaluation criteria.

Table 1 Model Performance Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | R2 |
| Random Forest | 12.39 | 207.201 | 14.394 | -0.006 |
| XGBoost | 13.094 | 243.379 | 15.601 | -0.182 |
| EBM | 12.398 | 205.422 | 14.333 | 0.002 |
| TabNet | 12.509 | 211.525 | 14.544 | -0.027 |



The Explainable Boosting Machine (EBM) emerged as the best-performing model, achieving the lowest RMSE of 14.333 and the highest R² value of 0.002, albeit modest. This superior performance aligns with EBM's design strengths in capturing complex non-linear relationships while maintaining interpretability. Random Forest demonstrated competitive performance with an RMSE of 14.394 and MAE of 12.39, showing robust predictive capability despite its ensemble nature. XGBoost showed the poorest results with the highest RMSE (15.601) and the most negative R² (-0.182). This suggests potential overfitting or suboptimal hyperparameter configuration for this specific epidemiological dataset. TabNet's performance fell in the middle range, with an RMSE of 14.544, indicating moderate effectiveness but not exceeding the simpler, more interpretable models.

The consistently low R² values across all models (ranging from -0.182 to 0.002) indicate that the linear variance explained by these models is limited. This suggests that Cholera incidence may be influenced by complex interactions, temporal dynamics, or unmeasured confounders not captured in the current feature set. However, the relatively similar MAE and RMSE values across the top-performing models (EBM, Random Forest, TabNet) suggest consistent predictive accuracy despite low explained variance. Given EBM's combination of superior performance and inherent interpretability, it was selected as the primary model for subsequent 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 2 presents the consolidated feature importance scores for the top 8 variables, ranked by their average importance across all models.

Table 2 Top 8 Feature Importance Rankings Across Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **Random Forest** | **XGBoost** | **EBM** | **TabNet** | **Average** |
| Infant Mortality Rate | 0.054 | 0.030 | 0.004 | 0.145 | 0.058 |
| Turbidity (NTU) | 0.048 | 0.021 | 0.013 | 0.094 | 0.044 |
| Water Treatment Method\_Filtration | 0.005 | 0.0184 | 0.000 | 0.0129 | 0.038 |
| Water Source Type\_Spring | 0.005 | 0.040 | 0.019 | 0.080 | 0.036 |
| Country\_Indonesia | 0.003 | 0.017 | 0.004 | 0.111 | 0.0034 |
| Water Source Type\_Well | 0.004 | 0.018 | 0.013 | 0.095 | 0.033 |
| Healthcare Access Index (0 - 100) | 0.050 | 0.027 | 0.010 | 0.027 | 0.029 |
| Contaminant Level (ppm) | 0.054 | 0.020 | 0.011 | 0.025 | 0.027 |

The analysis reveals that Infant Mortality Rate emerges as the most consistently important predictor across models, with an average importance score of 0.058. This socio-economic indicator likely serves as a proxy for broader healthcare infrastructure quality and population vulnerability. Turbidity, a direct water quality measure, ranks second with an average importance of 0.044, reflecting the critical role of water clarity and contamination in Cholera transmission. Notably, the feature importance values are consistently low across all models and features, with the highest individual score being 0.145 (TabNet for Infant Mortality Rate) and most values falling below 0.050. These low importance scores suggest that Cholera risk is driven by a distributed network of factors rather than a few dominant predictors, consistent with the multifactorial nature of waterborne disease transmission.

Additionally, we performed detailed feature importance analysis for each of the four models individually, examining the top 10 predictors for each approach (Figures A-D). These model-specific analyses reveal distinct patterns in feature selection and weighting strategies.

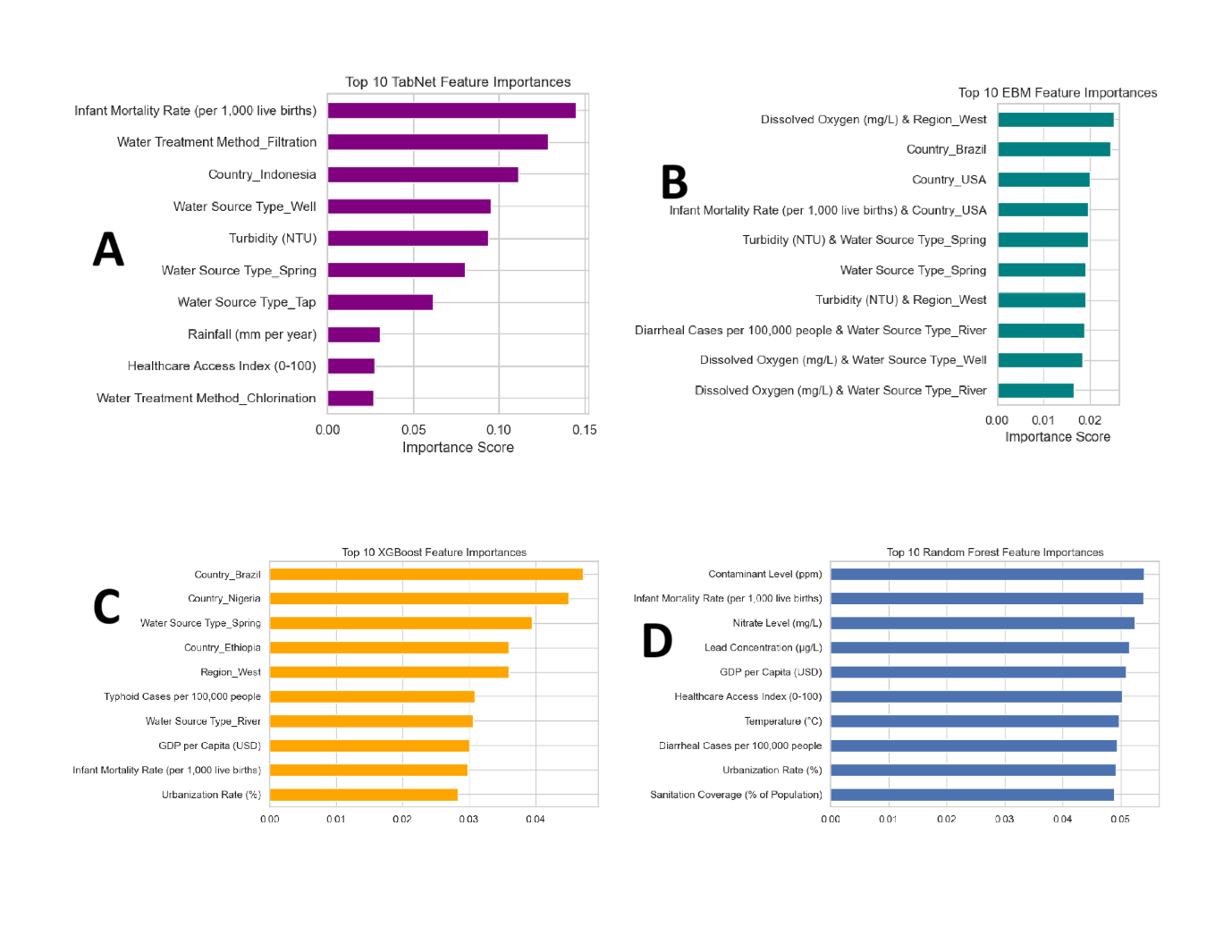


Figure A (TabNet) shows Infant Mortality Rate as the dominant predictor (importance ~0.145), followed by Water Treatment Method\_Filtration (~0.125) and Country\_Indonesia (~0.110). TabNet's attention mechanism appears particularly sensitive to socio-economic indicators and categorical variables related to water infrastructure and geographic location.

Figure B (EBM) demonstrates a more distributed importance pattern, with Dissolved Oxygen & Region\_West leading (~0.023), followed by various country-specific and water source interactions. EBM's focus on interaction terms (e.g., "Dissolved Oxygen & Water Source Type\_Well") reflects its ability to capture feature combinations, though all scores remain below 0.025.

Figure C (XGBoost) prioritizes geographic factors, with Country\_Brazil and Country\_Nigeria showing the highest importance (~0.042-0.045), followed by water source types and regional indicators. This geographic emphasis suggests XGBoost may be capturing country-specific epidemiological patterns and regional clustering effects.

Figure D (Random Forest) emphasizes direct environmental and health indicators, with Contaminant Level, Infant Mortality Rate, and Nitrate Level ranking highest (all ~0.048-0.055). The prominence of chemical contaminants (nitrate, lead) and sanitation coverage aligns with established Cholera transmission pathways.

## **6.3 Statistical Validation of Key Predictors**

### **6.3.1 Poisson Regression Analysis**

To complement the machine learning models and statistical correlation analysis, we conducted a Poisson regression analysis to further investigate the association between predictor variables and the incidence of Cholera. Poisson regression is particularly suitable for modeling count-based outcome variables, such as disease incidence rates, and allows for direct estimation of incidence rate ratios (IRRs) while controlling for multiple covariates simultaneously.

This approach enabled us to quantify the relative contribution of key environmental, socio-economic, and infrastructural predictors to Cholera cases per 100,000 people, while accounting for potential confounding effects. The model included all relevant predictors that passed prior significance thresholds in the correlation and regression analyses. Results from the Poisson regression are presented visually through three plots:

Figure 4 presents the results of the Poisson regression analysis using three complementary visualizations, each offering key insights into the environmental and socio-economic factors associated with cholera incidence. The first plot (A), a forest plot, reveals several variables with statistically significant positive associations with cholera cases. Among the most notable are the pH level of water, contaminant level (ppm), and lead concentration (µg/L), all of which point to water quality as a critical driver of cholera risk. Diarrheal cases per 100,000 people also emerge as a strong predictor, likely indicating overlapping infrastructural vulnerabilities that facilitate both cholera and other waterborne diseases. Additionally, higher ambient temperatures and elevated dissolved oxygen levels are associated with increased cholera incidence, aligning with the known environmental preferences of Vibrio cholerae.

The second plot (B), displaying incidence rate ratios (IRRs), confirms that increases in factors such as pH, turbidity, contaminant level, and temperature are consistently linked to elevated cholera risk. Though the IRRs are modest in magnitude, their statistical significance underscores the public health importance of these variables. Interestingly, urbanization rate also shows a slight positive association, suggesting that rapid urban growth—often accompanied by inadequate infrastructure—may contribute to increased disease vulnerability.

The third plot (C), a volcano plot, compares the statistical and practical significance of each variable's effect on cholera cases. While many variables are statistically significant, their effect sizes tend to be modest, falling within a narrow practical range. This indicates that cholera risk is not driven by a single dominant factor but rather by a network of interrelated influences. No variables exhibit both high statistical and practical significance, reinforcing the multifactorial nature of cholera transmission. Overall, these findings emphasize the need for integrated, multi-dimensional public health interventions that address a broad set of environmental and infrastructural conditions to effectively mitigate cholera risk.

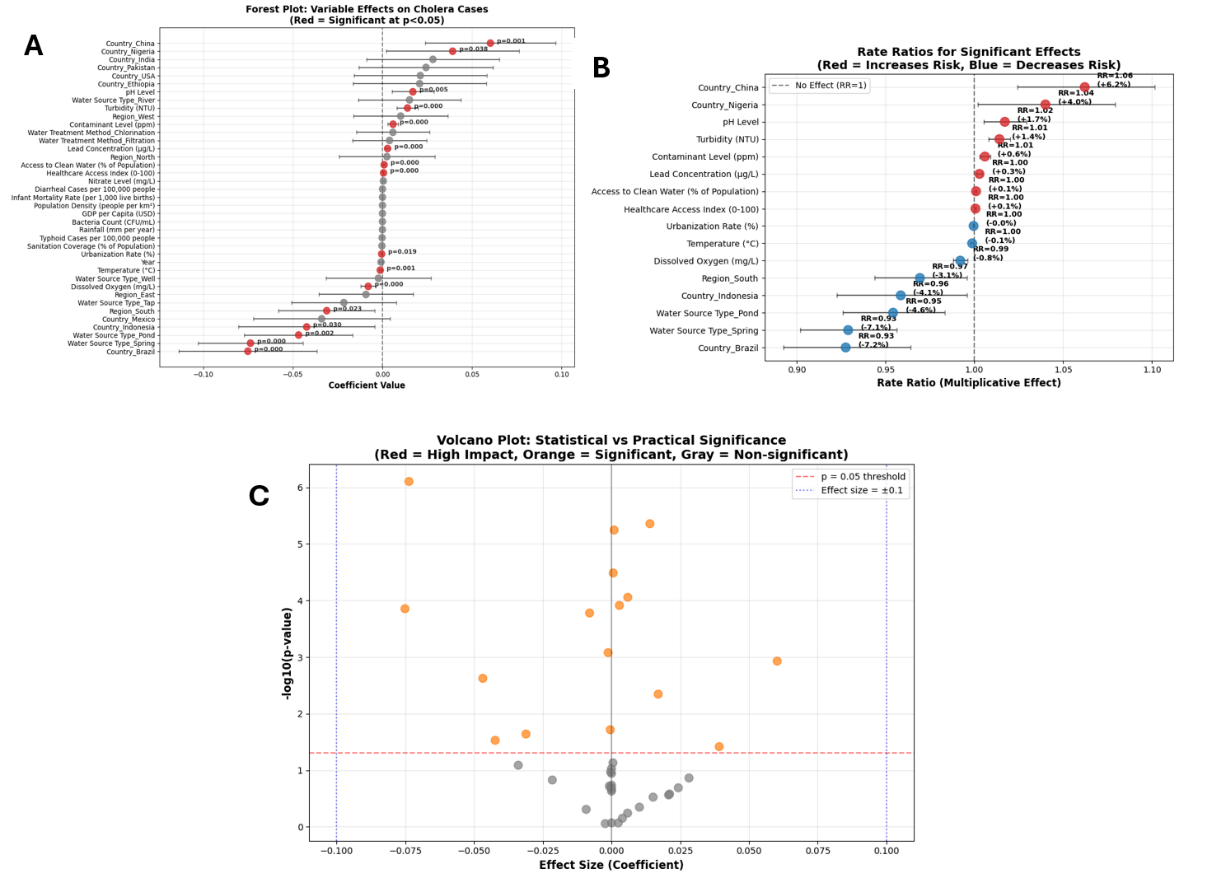
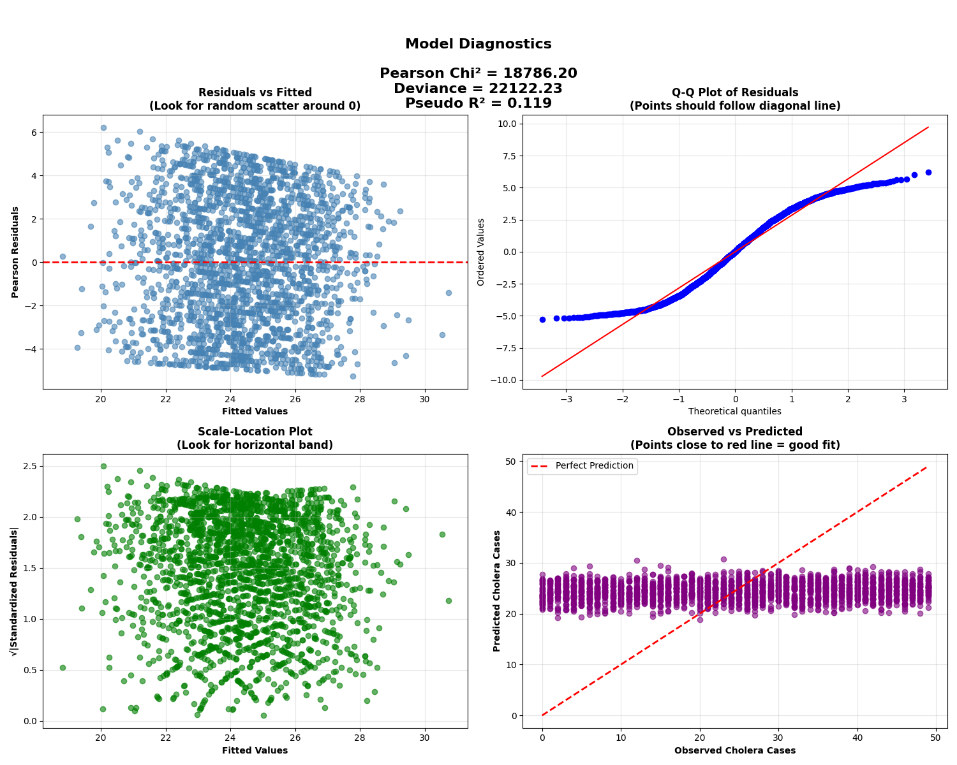


Figure 4 Poisson regression analysis

Figure 5 below presents the model diagnostic plots for the Poisson regression predicting cholera cases, offering a comprehensive evaluation of model fit and underlying assumptions. The Pearson residuals vs. fitted values plot (top-left) is expected to show a random scatter of points around zero. However, the observed dispersion and slight funnel shape suggest potential heteroscedasticity, indicating that residual variance may increase with higher fitted values. This challenges the assumption of constant variance and suggests possible under- or overestimation in specific ranges. The Q-Q plot (top-right), which evaluates the normality of residuals, shows notable deviations from the diagonal red line, particularly in the tails. Although strict normality isn't required for Poisson models, this deviation may reflect unmodeled complexity, outliers, or limitations in the log-linear structure of the model. The scale-location plot (bottom-left), ideally displaying a horizontal spread of points, shows some variance across fitted values. This further supports concerns about heteroscedasticity and suggests that predictive performance may not be consistent across the outcome range.

The observed vs. predicted plot (bottom-right) compares actual and predicted cholera case counts. The clustering of points away from the 45-degree red reference line indicates systematic bias, with the model failing to capture the full variability in observed data. This is consistent with the model’s pseudo R² of 0.18, suggesting limited explanatory power. While the Poisson model identifies significant associations, Figure 5 highlights concerns related to residual behaviour, prediction accuracy, and variance explanation. These findings indicate the potential need for model enhancements, such as incorporating interaction terms, exploring nonlinear effects, or adopting alternative approaches like negative binomial regression to better account for overdispersion and improve overall model fit.



## **6.4 Demonstration of Agent Response and Interface Functionality**

### **6.4.1 Prompt Engineering for the Agent**

To ensure accurate and contextually relevant responses, the agent is guided by a carefully constructed prompt template developed using the PromptTemplate module from langchain\_core.prompts. This prompt defines the agent’s role, access to data, analytical approach, and communication style, forming the foundation for consistent and reliable natural language interactions. The agent is explicitly instructed to operate as a Cholera Risk Analysis Agent, capable of interpreting complex epidemiological and environmental data using outputs from statistical analyses and machine learning models.

The prompt embeds five structured datasets into the agent’s context: (1) Feature Importance, which includes variable rankings from EBM, NGBoost, and TabNet models along with aggregated importance scores; (2) Pearson Correlation, capturing linear relationships between predictors and cholera incidence; (3) Spearman Correlation, which identifies monotonic, potentially non-linear associations and is robust to outliers; (4) Linear Regression, providing coefficients and corresponding p-values to assess statistical significance in a multivariate setting; and (5) Model Performance, summarizing predictive accuracy via RMSE scores across the three models.

To maintain analytical rigor, the prompt enforces several best practices. The agent is required to ground its answers in actual data values—citing specific correlation coefficients, model rankings, and regression outputs where relevant. It must also interpret statistical metrics in plain language, ensuring clarity for non-specialist users. Furthermore, it emphasizes the distinction between correlation, regression, and feature importance to prevent misinterpretation. Finally, all responses are framed within a public health context, offering actionable insights such as improvements in sanitation or water safety based on the findings.

This structured prompting strategy enables the agent to deliver transparent, traceable, and evidence-based insights. It ensures that the system not only answers user queries accurately but also supports informed decision-making in public health—where clarity, validity, and accountability are paramount.

### **6.4.2 Interactive Interface**

This section presents selected outputs from the deployed Streamlit-based agent interface described in Section 5. These results illustrate the system’s core capabilities—namely, generating interpretable, evidence-based responses to user queries grounded in structured data analysis. We demonstrate how the intelligent agent handles a variety of question types, including identifying key predictors of Cholera incidence, comparing model performance, and recommending targeted interventions. Through these examples, we showcase the agent’s ability to integrate natural language processing, machine learning outputs, and statistical reasoning in a cohesive, user-friendly interface. Each response is drawn from actual data and models, derived from the implementation as detailed in section 5.0.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure 3 Deployed Q&A interface showing traceable, data-driven responses to Cholera-related queries, with citations from correlation and feature importance analyses.

**Figure 3** presents a screenshot of the deployed Q&A interface, illustrating the interaction between user queries and LLM-generated responses grounded in structured data analysis. The interface displays structured data citations from correlation matrices and feature importance analyses, demonstrating the system's ability to provide traceable, data-driven epidemiological insights.

### **6.4.3 Visualizing 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:**

A screenshot of a computer

AI-generated content may be incorrect.

Upon receiving the user query, the agent immediately identified and corrected a factual inaccuracy in the question—specifically, the claim that there was a strong positive correlation (+0.75) between water contamination and cholera incidence. Instead, it accurately reported that the Spearman correlation between *Contaminant Level (ppm)* and *Cholera Cases per 100,000 people* is only 0.028, indicating a very weak positive relationship. This correction demonstrates the agent’s ability to cross-check user input against the actual data and provide scientifically valid clarification—an essential feature for maintaining the integrity of public health insights.

The agent further validated its response using multiple layers of evidence:

* 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.

# **7.0 Discussion**

The research question for this study was: ***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 how such an integrated framework can be achieved and operationalized, providing a comprehensive answer to this core research question. The integration was achieved through a multi-layered approach that combines complementary analytical methodologies with intelligent natural language interpretation, creating a comprehensive system that bridges the gap between complex epidemiological analysis and accessible decision support.

## **7.1 Framework Innovation and Methodological Advances**

The framework's success in answering this research question was achieved through several key innovations and methodological advances. First, the development of a multi-model analytical pipeline leveraging Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet demonstrated how diverse interpretable machine learning approaches could be systematically combined to analyze Cholera risk factors. EBM's glass-box interpretability provided transparent feature importance rankings and interaction visualizations, while TabNet's attention mechanism revealed complex categorical relationships particularly sensitive to socio-economic indicators and water infrastructure variables. The inclusion of Random Forest and XGBoost as benchmark models ensured robust comparative analysis, ultimately revealing that EBM achieved superior performance with the lowest RMSE (14.333) while maintaining full interpretability—a critical requirement for public health applications.

Second, the integration of comprehensive statistical validation methods provided essential verification of machine learning findings through traditional epidemiological approaches. Pearson and Spearman correlation analyses captured both linear and monotonic relationships between predictors and Cholera incidence, while multivariate linear regression controlled for confounding effects and assessed statistical significance. The Poisson regression analysis was particularly valuable for modelling count-based disease outcomes, revealing that water quality indicators such as pH, contaminant levels, and lead concentrations showed statistically significant associations with Cholera cases. This multi-method statistical approach ensured that machine learning-identified predictors were not merely algorithmic artefacts but represented genuine epidemiological relationships with demonstrated statistical validity.

## **7.2 LLM-Powered Natural Language Decision Support**

Third, the development of a LangChain-powered LLM agent that interprets model outputs and statistical analyses enabled natural language question-answering for decision support represented a critical innovation in making complex epidemiological insights accessible. The agent's sophisticated prompt engineering embedded structured CSV outputs from all analytical components—feature importance rankings, correlation matrices, regression coefficients, and model performance metrics—directly into the LLM's context. This approach ensured that every response was grounded in actual data rather than speculative reasoning, as demonstrated when the agent correctly identified and corrected factual inaccuracies in user queries while providing evidence-based recommendations. The Streamlit interface successfully democratized access to complex analytical results, allowing public health practitioners and researchers to explore data-driven relationships through intuitive natural language interactions.

**7.3 Consistent Risk Factor Identification and Epidemiological Insights**

Fourth, the framework's ability to identify consistent key risk factors across multiple regions was evidenced through convergent findings across all analytical approaches. Infant Mortality Rate emerged as the most important predictor across models (average importance 0.058), serving as a proxy for broader healthcare infrastructure quality and population vulnerability. Water quality indicators, particularly turbidity (0.044 average importance) and contamination levels, consistently ranked among top predictors, validating established understanding of waterborne disease transmission pathways. Geographic factors such as country-specific effects (Indonesia, Brazil, Nigeria) highlighted regional epidemiological variations, while water infrastructure variables (treatment methods, source types) underscored the critical role of WASH interventions. Importantly, the consistently low feature importance values and modest correlation coefficients across all methods revealed that Cholera risk emerges from complex, multifactorial interactions rather than single dominant predictors—a nuanced epidemiological insight that single-method analyses might have missed.

## **7.3 Paradigmatic Contribution to Epidemiological Research**

The framework's success in integrating explainable machine learning with statistical reasoning through LLM-powered interpretation demonstrates a novel paradigm for epidemiological research and decision support. By combining the pattern recognition capabilities of modern machine learning with the inferential rigour of statistical analysis and the interpretive intelligence of large language models, this approach offers a replicable methodology for addressing complex public health challenges that require both analytical sophistication and practical accessibility.

This integrated approach addresses fundamental challenges in contemporary epidemiological research, where the growing availability of diverse data sources—including clinical records, environmental monitoring, social media, and administrative databases—demands analytical frameworks capable of processing multimodal information while maintaining interpretability for clinical decision-making. The framework's modular design, combining specialized analytical components with intelligent orchestration, represents a significant advancement over traditional single-method approaches that often fail to capture the multifactorial nature of disease risk.

## **7.4 Methodological Validation and Robustness**

The comprehensive validation strategy employed in this study—incorporating both internal cross-validation and external consistency checks across multiple analytical methods—ensures the robustness and reliability of findings. The convergent evidence from diverse methodological approaches (interpretable machine learning, classical statistics, and LLM-powered synthesis) provides strong support for the identified risk factors while maintaining the transparency essential for public health decision-making. This multi-method validation approach addresses a critical limitation in many epidemiological studies where single analytical frameworks may introduce method-specific biases or miss important relationships. The framework's ability to automatically generate structured outputs suitable for LLM interpretation while preserving the statistical rigour of traditional epidemiological methods represents a significant technical achievement. This capability enables the seamless integration of quantitative analytical results with qualitative interpretation, creating a system that can adapt to diverse user needs and expertise levels without compromising analytical integrity.

In conclusion, this study demonstrates that LLM-powered agentic frameworks can effectively integrate explainable machine learning and statistical reasoning to identify key risk factors for Cholera incidence across multiple regions. The success of this integration lies not merely in the technical combination of different analytical approaches, but in the intelligent orchestration of these methods through natural language interfaces that make sophisticated epidemiological insights accessible to decision-makers. This paradigm offers a promising foundation for addressing complex public health challenges in an era of increasing data availability and analytical sophistication, while maintaining the interpretability and trustworthiness essential for evidence-based public health practice.

# **8.0 Conclusion and Future Work**

Key findings revealed that Cholera risk emerges from complex, multifactorial interactions rather than single dominant predictors, with infant mortality rate serving as the most consistent socio-economic indicator and water quality measures (turbidity, contamination levels) representing critical environmental drivers. The EBM model achieved superior predictive performance (RMSE 14.333) while maintaining full interpretability, validating the value of explainable AI in epidemiological applications. The LLM agent successfully democratized access to sophisticated analytical results, enabling natural language exploration of complex epidemiological relationships while maintaining scientific rigor through data-grounded responses. This work demonstrates the transformative potential of combining explainable AI with LLM-powered interpretation for public health analytics, offering a replicable framework for evidence-based disease prevention strategies in resource-limited settings where timely, interpretable guidance is critical.

Several promising directions emerge for extending this framework's capabilities and impact. First, temporal modeling could be incorporated to capture seasonal patterns and outbreak dynamics by integrating time-series analysis with the existing spatial risk factor identification. This would enable early warning systems and dynamic risk assessment as environmental and socio-economic conditions change over time. The framework could be expanded to include additional waterborne diseases (dengue, typhoid, hepatitis) to develop a comprehensive WASH-related disease surveillance system. This multi-disease approach would identify shared risk factors and enable integrated intervention strategies that address multiple health threats simultaneously. Real-time data integration capabilities could transform the system into a live monitoring platform by incorporating APIs for weather data, water quality sensors, and health surveillance systems. This would enable continuous risk assessment and automated alert generation for public health authorities.

Fourth, the LLM agent could be enhanced with multi-modal capabilities to process satellite imagery, clinical photographs, and environmental sensor data alongside tabular analytics. This would provide richer contextual understanding and support field-based decision-making through mobile applications. Lastly, deployment studies in low-resource settings would validate the framework's practical utility and cultural appropriateness while identifying implementation challenges and adaptation requirements. Partnerships with ministries of health and international organizations could facilitate pilot implementations and impact evaluation, ultimately advancing the translation of AI-powered epidemiological tools into routine public health practice.

# **9.0 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. Olayinka, H.M.: Developing Explainable Machine Learning Models for Early Diagnosis, Prognosis, and Personalized Treatment Planning in Complex Chronic Diseases: A Comprehensive Study on Data Integration, Ethical Challenges, and Clinical Deployment. Int. J. Future Eng. Innov. 2, 113–120 (2025). https://doi.org/10.54660/IJFEI.2025.2.3.113-120.

4. Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., He, X.: Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects, http://arxiv.org/abs/2401.03428, (2024). https://doi.org/10.48550/arXiv.2401.03428.

5. 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.

6. 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.

7. Charnley, G.E.C., Kelman, I., Green, N., Hinsley, W., Gaythorpe, K.A.M., Murray, K.A.: Exploring relationships between drought and epidemic cholera in Africa using generalised linear models. BMC Infect. Dis. 21, 1177 (2021). https://doi.org/10.1186/s12879-021-06856-4.

8. 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).

9. 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.

10. 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.

11. 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.

12. 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.