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

Paul Jideani

Aurona Gerber

SACAIR2025

UNIVERSITY of the WESTERN CAPE

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

Table of Contents

01. The Big Picture

02. Introduction

03. Research Question

04. Dataset
- Description
- Preprocessing

05. Methodology
- Knowledge base
- LLM framework

06. Langchain + Langraph Integration

07. Streamlit UI

08. Results
- Evaluation – Model Performance
- Feature Importance
- Model Performance

09. Example – User query

10. The Interface

11. Sample Agent response

12. Discussion

13. Conclusion

The Big Picture

Cholera is a waterborne disease and a major public health challenge.

To predict outbreaks or analyse cholera patterns, we ideally need reliable data such as water quality reports, water treatment records, demographic information, and epidemiological or clinical data.

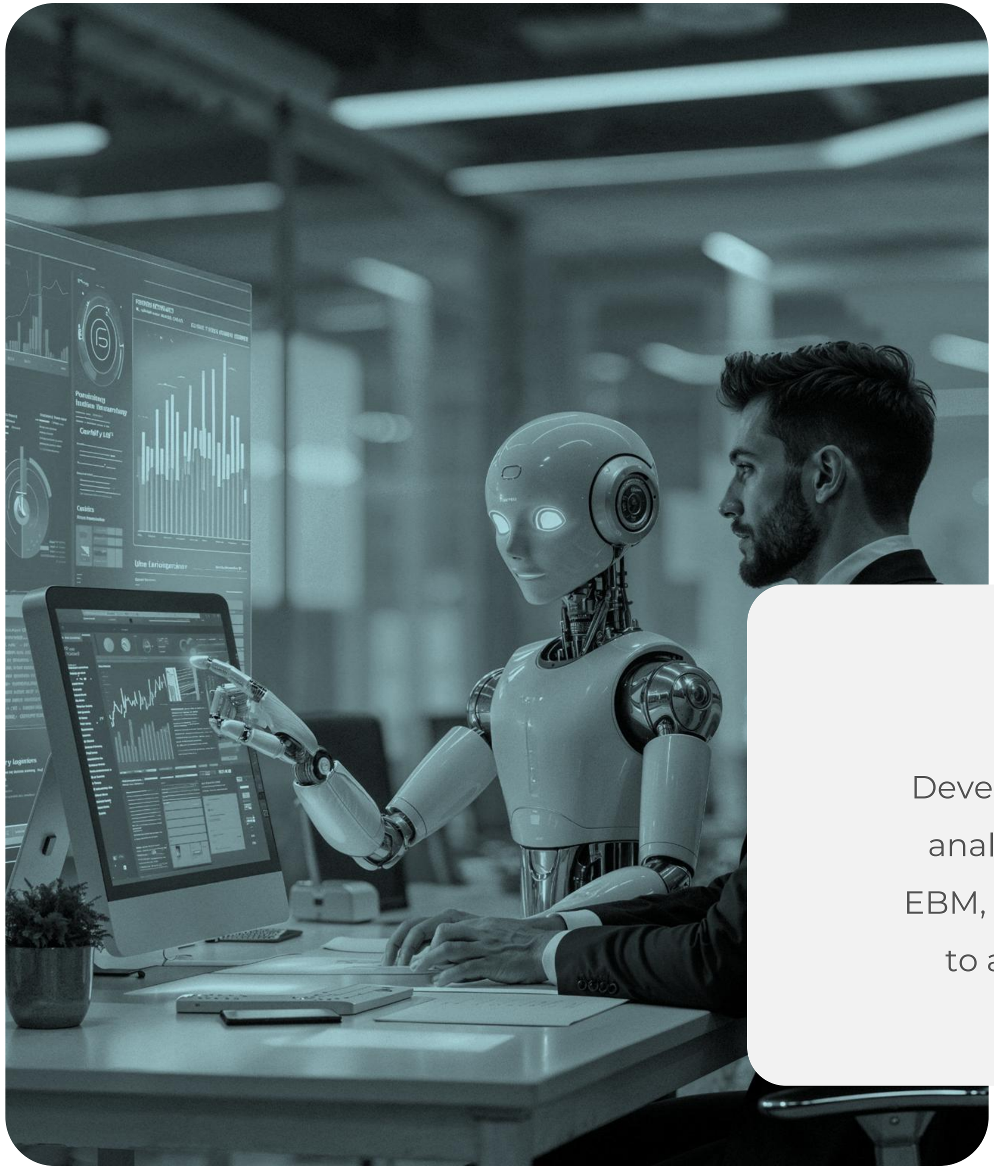
But in many instances, these types of data are limited or unavailable.

The question then becomes

What can we do in a data-constrained environment?

When traditional data sources are missing, we can turn to alternative forms of information. Social media posts, publicly available datasets, environmental data, mobility patterns, and other unconventional sources can help fill the gaps.

By combining these alternative data streams, we can still perform meaningful analyses and predictions of cholera. Using non-traditional, accessible data to overcome resource limitations is the core of my research.



Introduction

With the big picture in mind, this study *narrowly* examined how interpretable machine learning and large language model (LLM)–driven reasoning can collaborate to provide transparent and actionable insights.

Specifically, the research investigates how an LLM-powered agentic framework can integrate explainable ML models with statistical reasoning to identify key risk factors for cholera across multiple regions.

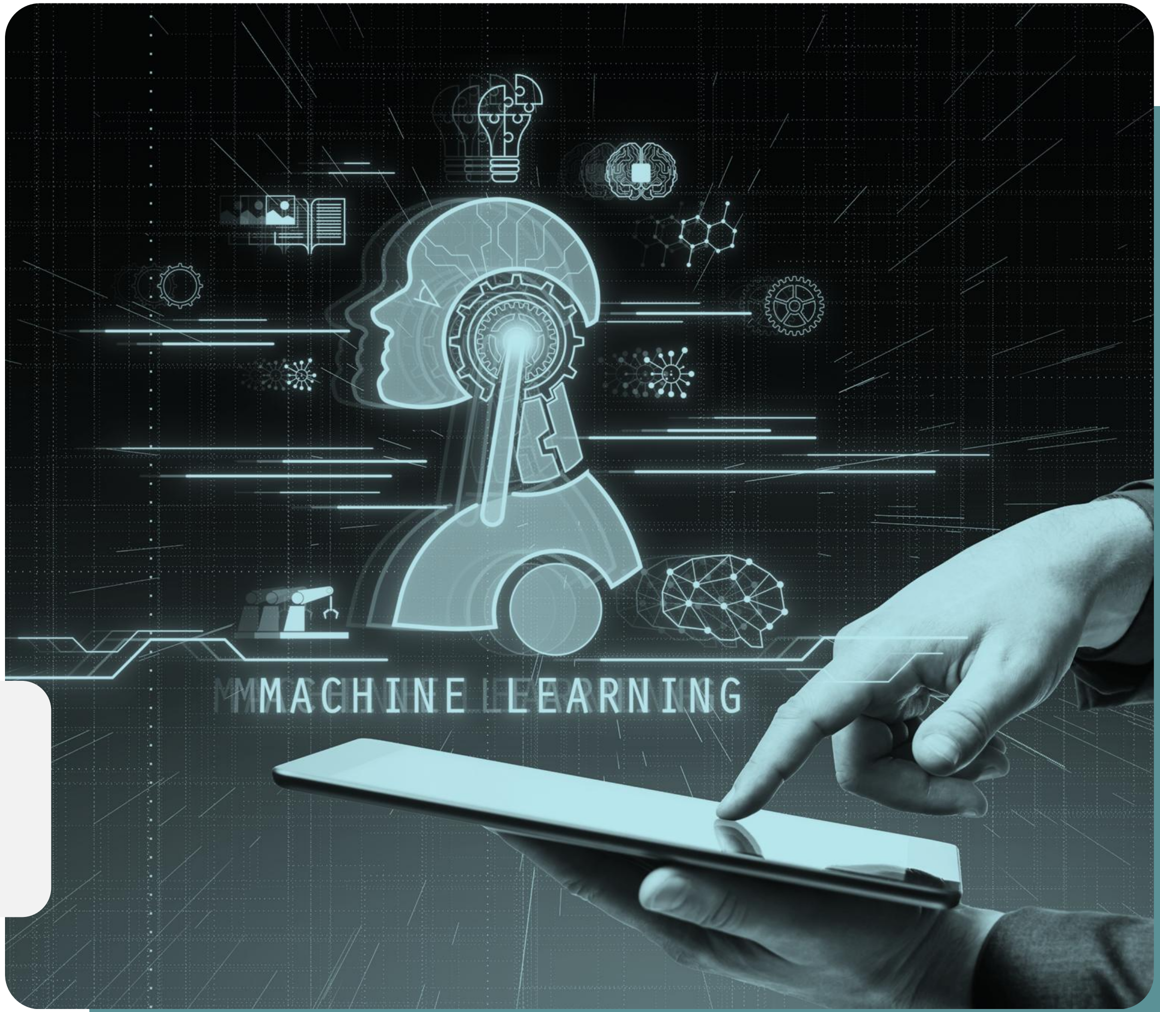
This work contributes by

Developing a multi-model analytical pipeline using EBM, NGBoost, and TabNet to analyse cholera risk factors.

Integrating a LangChain-based LLM agent that interprets model outputs, supports natural language querying, and enhances decision-making in data-constrained environments.

Research Question

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?



Bottom line

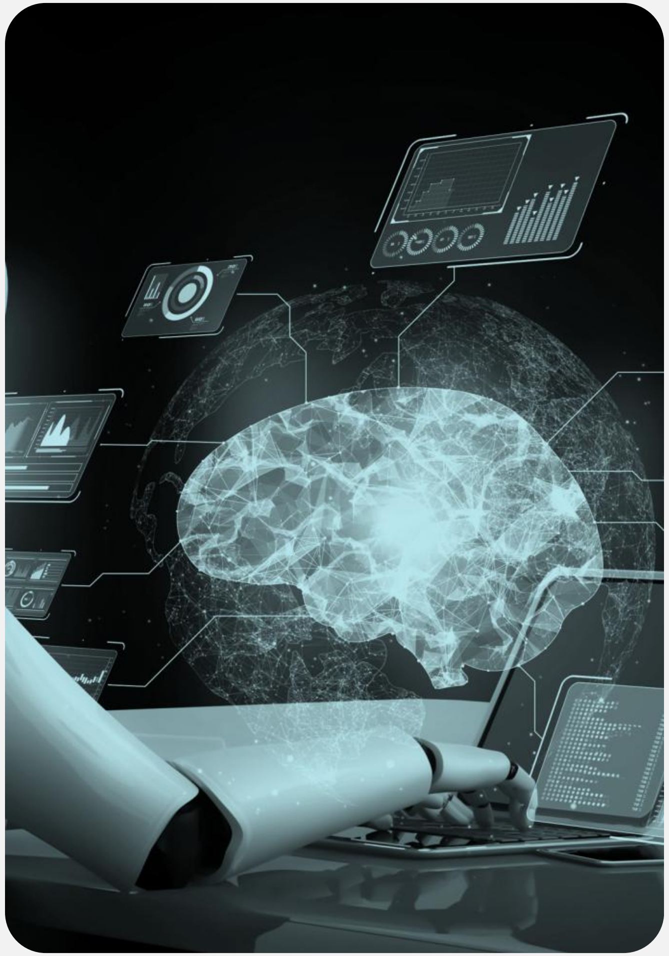
When ideal data isn't available, we use explainable AI and natural language processing to extract insights from what we do have, making cholera prediction or analyses possible.

Dataset Description

Overview

Provides a contemporary and comprehensive view of water pollution indicators and public health outcomes.

Includes environmental metrics, socio-economic conditions, and cholera incidence across multiple regions.



Source

Water Pollution and Disease dataset from Kaggle

Published: **March 2025**

Coverage: **2000 – 2025**

Focus: Environmental, socio-economic, and health factors influencing waterborne diseases, especially **Cholera**

...Dataset Description

Variables

Target Variable

Cholera Cases per 100,000 people

Predictor Variables

Environmental

Rainfall frequency

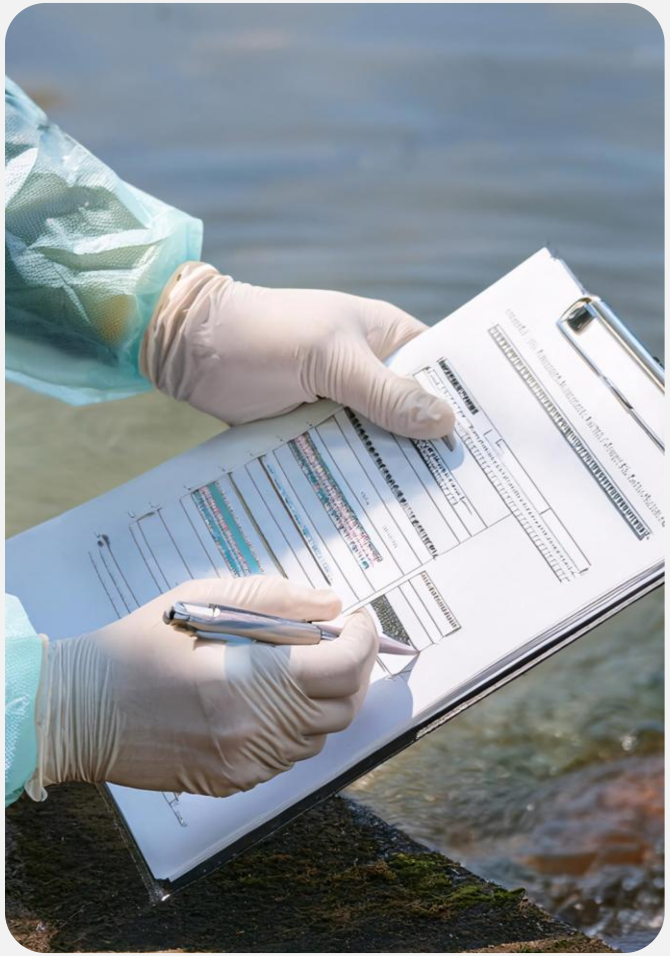
Water source type

Contaminant levels

pH, turbidity, dissolved oxygen

Nitrate and lead concentrations

Bacterial counts



Predictor Variables

Socio-economic

Sanitation access

Open defecation rates

Education levels

Population density

Infrastructure-related variables

Dataset Pre-processing

01.

02.

03.

04.

05.

Target Variable Validation

- Confirmed the presence of Cholera Cases per 100,000 people before transformation.

Feature Categorisation

- Separated numerical and categorical features.
- Target variable excluded from all preprocessing steps.

Skewness Handling

- Calculated skewness coefficients for numerical features.
- Applied log1p transformation to variables with skew > 1 to reduce outlier impact and stabilise feature distributions.

Normalisation

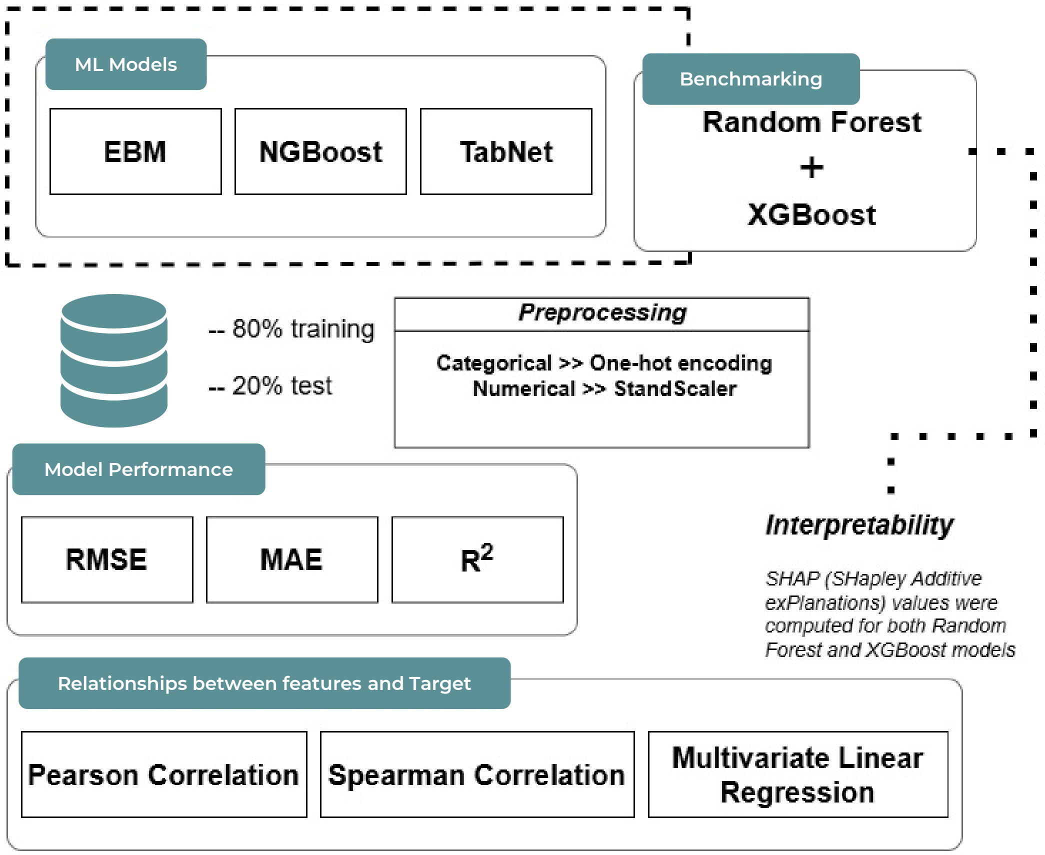
- Standardised numerical features using z-score normalisation for consistent scaling across all variables.

Categorical Encoding

- Used one-hot encoding with the first category dropped to avoid multicollinearity.

Methodology – Knowledge base

Methodology



Methodology – LLM Framework

Data Sources

- pearson_corr-matrix.csV
- spearman_corr_matrix.csV
- feature_importance_comparison.csv
- model_performance_comparsion.csv
- linear_regression_coefficients.csv

CSV Data Ingest

LangChain + LangGraph Framework

Memory Saver
Context Tracking

Data Loader
CSV Processing

Custom Prompt Template
Structured data injection:

LangGraph Agent Node
1. Load CSVs as formatted strings
2. Populate QA prompt with structured data
3. Invoke Cluade LLM for response generation

Claude LLM
Evidence-grounded response generation

Multi-turn Dialogue Support
Conversation context & follow-up handling

Grounded Response

User Query

LLM Response

Streamlit User Interface

Chatbot-style Interface

User: "What are key Cholera predictors?"

Assistant: "Based on feature importance analysis
Water Access Index (0.847), Population
Density (0.623)
Source:
feature_importance_comparison.csv

Free-form Input Box
"Ask me a question about Cholera"

System Capabilities
- Evidence-traceable answers
- Natural language interaction
- Rapid data exploration
- Policy-relevant recommendations
- Real-time loading feedback

Langchain + Langraph Integration

Framework Structure

- Built using LangChain's LangGraph to enable stateful, multi-step reasoning through a directed graph.
- Implemented as a minimal LangGraph with a single question-answering node supported by structured memory.

Data-Grounded Reasoning

- The agent loads outputs from ML and statistical analyses (correlations, coefficients, importance rankings).
- Each CSV is formatted and injected into a custom prompt template to ensure evidence-based responses.

Prompt Design

- Delimited sections for

Correlation matrices | Regression coefficients | Feature importance rankings

- Enables precise, context-aware answer generation.

Execution Logic

- Load and format the required CSV files
- Populate QA prompt with structured analytics
- Append user query and history
- Invoke Claude LLM for grounded response generation

Conversation Memory

- MemorySaver ensures multi-turn dialogues remain coherent, allowing the agent to track context across interactions.

Streamlit UI

Interaction Flow

01. User enters a question

02. Message is added to the conversation state

03. Query is sent to the **LangGraph agent**

04. The LLM response is appended and displayed in the chat interface

Lightweight Web Interface

- Built using **Streamlit** to make the framework accessible to non-technical users.

Key Features

- Chatbot-style layout with persistent conversation history
- Free-form input box for user queries
  - e.g., “What are the most important features for predicting Cholera?”
- Dynamic loading spinner for real-time response feedback

Evaluation – Model Performance

We employed four commonly used regression metrics to provide predictive accuracy and model fit.

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

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

Results – Feature Importance

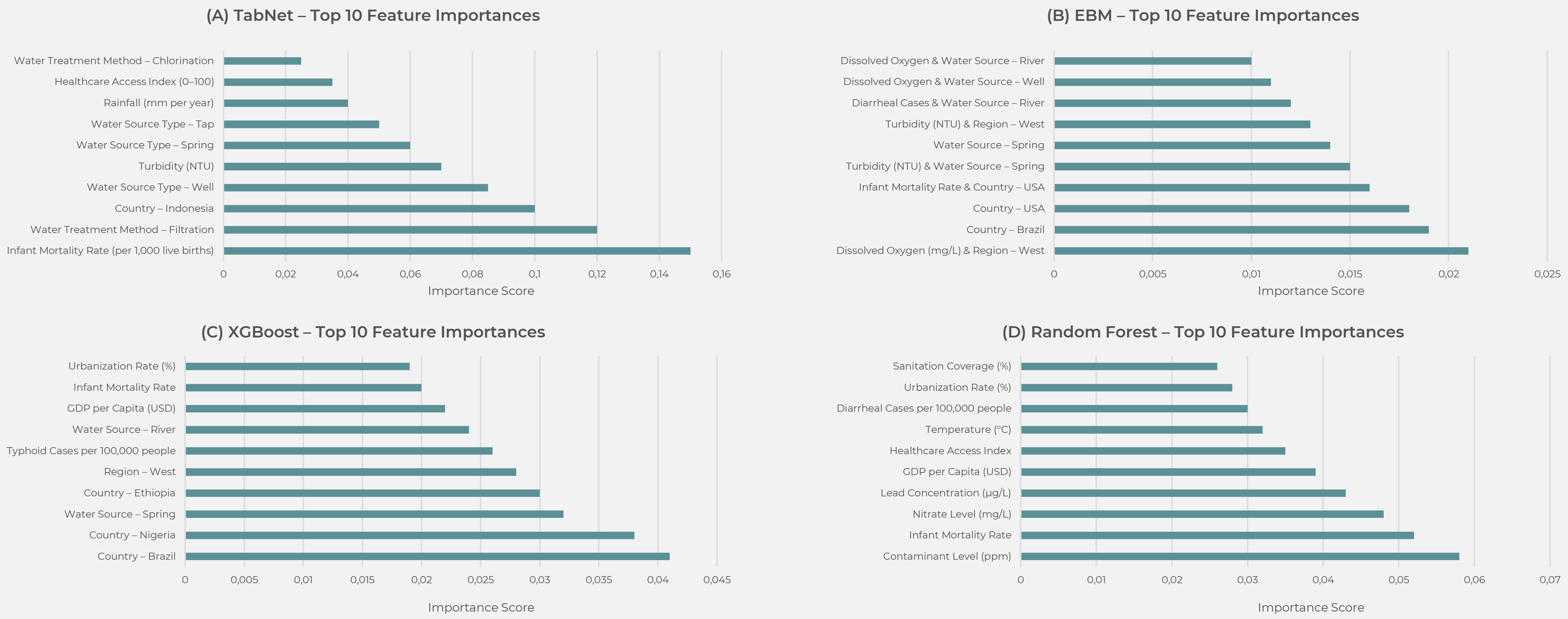
The analysis reveals that Water Contaminant Level (ppm) and Water Treatment Method (Filtration) are the strongest predictors of cholera incidence, particularly in Random Forest (1.54/1.50) and TabNet (1.250/1.129), indicating that chemical contamination and inadequate treatment are significant drivers of infection. The Infant Mortality Rate also stands out as a consistent indicator across models (0.540, 0.400, 0.300, 0.145), reflecting the underlying vulnerability of the health system. The Water Source Type (Spring) (0.800 / 0.500) further highlights the risks associated with untreated water sources, while Turbidity shows a moderate but meaningful influence (0.480 vs. 0.094). In contrast, the Healthcare Access Index displays lower importance (0.300, 0.170, 0.400, 0.111), suggesting that although healthcare access matters, water quality factors play a more dominant role in predicting cholera transmission.

Feature	Random Forest	XGBoost	EGM	TabNet
Infant Mortality Rate	**0.540**	**0.300**	**0.40**	**0.145**
Turbidity (NTU)	**0.480**	**0.021**	**0.013**	**0.094**
Water Treatment Method_Filtration	**1.50**	**1.184**	**0.000**	**1.129**
Water Source Type_Spring	**0.50**	**0.400**	**0.190**	**0.800**
Healthcare Access Index (0 - 100)	**0.300**	**0.170**	**0.400**	**0.111**
Contaminant Level (ppm)	**1.54**	**1.200**	**0.110**	**1.250**

Results – Feature Importance

Figure 4 and subfigures A – D show model-specific top-10 predictors,
highlighting distinct feature selection and weighting patterns across the four models.

Top 10 Feature Importances Across Models



Results – Model Performance

We employed four commonly used regression metrics to provide predictive accuracy and model fit.

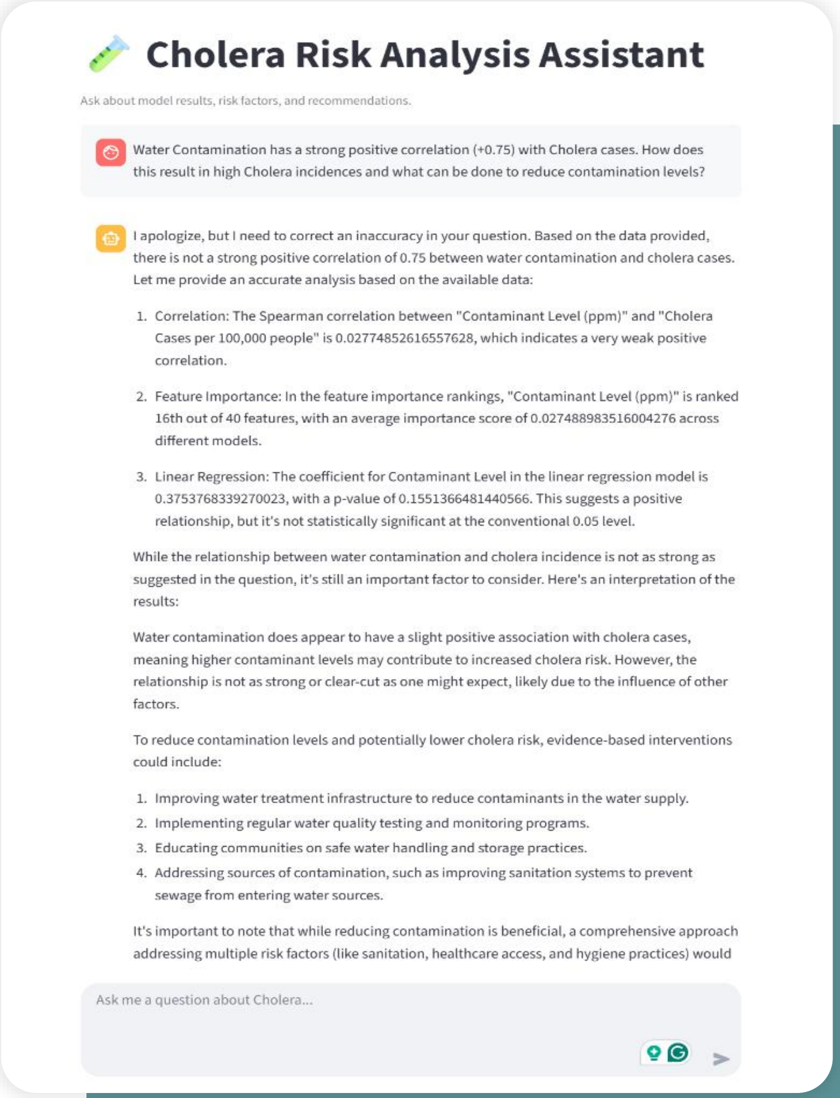
Model	MAE	MSE	RMSE	R^2
Random Forest	12.39	207.201	14.394	-0.006
XGBoost	13.094	243.379	15.601	-0.182
EGM	12.398	205.422	14.333	0.002
TabNet	12.509	211.525	14.544	-0.027

Example - User query

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?”

The Interface



Example - Agent Response

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:

01.

Correlation

The Spearman correlation between "Contaminant Level (ppm)" and "Cholera Cases per 100,000 people" is 0.0277, which indicates a very weak positive correlation.

02.

Feature Importance

In the feature importance rankings, "Contaminant Level (ppm)" is ranked 18th out of 40 features, with an average importance score of 0.027 across different models.

03.

Linear Regression

The coefficient for Contaminant Level in the linear regression model is 0.3753, with a p-value of 0.155. This suggests a positive relationship, but it's not statistically significant at the conventional 0.05 level.

Example - Agent Response (Cont'd)

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.

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

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

01. Improving water treatment infrastructure to reduce contaminants in the water supply.

02. Implementing regular water quality testing and monitoring programs.

03. Educating communities on safe water handling and storage practices.

04. Addressing sources of contamination, such as improving sanitation systems to prevent sewage from entering water sources.

Conclusion

This framework offers an approach to epidemiological analysis by uniting the strengths of machine learning, statistical inference, and LLM-driven interpretation.

By blending ML's pattern-recognition abilities with the explanatory rigour of traditional statistical methods and the natural-language reasoning capabilities of LLMs, it delivers insights that are both analytically robust and highly accessible.

Its modular design ensures that complex analyses can be conducted transparently, enabling diverse stakeholders, such as policymakers and public health practitioners, to explore, question, and understand epidemiological patterns with ease.

With future improvement, this integrated approach enhances decision-making by making advanced analytical tools interpretable, trustworthy, and usable in real-world public health contexts, especially in settings with limited resources.

Thank You

Paul Jideani

Prof Aurona Gerber

SACAIR2025

UNIVERSITY of the WESTERN CAPE

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT