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