

PREDICTIVE AND PRESCRIPTIVE ANALYTICS TO STRENGTHEN INFECTIOUS DISEASE SURVEILLANCE AND RESPONSE

JOY FLORENCE AWOR

S23B38/022

CALVIN DIEGO RWOTHOMIO

S23B38/019

JUDITH ANITA NAMAGANDA

S23B38/025

**A PROJECT REPORT SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND
TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD
OF A DEGREE OF BACHELOR OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF
UGANDA CHRISTIAN UNIVERSITY**

June, 2026



**UGANDA CHRISTIAN
UNIVERSITY**

A Centre of Excellence in the Heart of Africa

Declaration

We, **Awor Joy Florence, Judith Anita Namaganda, Rwothomio Calvin Diego**, declare to the best of our knowledge that the work in this proposal is original and has not been submitted before to any university/institution. It is the result of our effort.

Awor Joy Florence

Signature:

Date:

Judith Anita Namaganda

Signature:

Date:

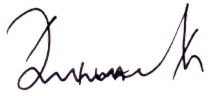
Rwothomio Calvin Diego

Signature:

Date:

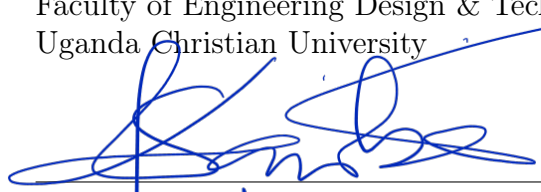
Approval

This proposal has been submitted for examination with the approval of the following advisors/supervisor.

Signature: _____

Date: 01/06/2026

Dr. Terhemba Micheal-Ahile
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature: _____

Date: 01/06/2026

Dr. Samuel Kakuba
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Abstract

Uganda continues to experience recurring outbreaks of epidemic-prone infectious diseases, driven by persistent gaps in water, sanitation, and hygiene infrastructure, rapid urbanization, population displacement, and a limited capacity for timely detection and coordinated response. Current surveillance systems remain largely reactive and descriptive, focusing on reporting confirmed cases rather than predicting and preventing outbreaks. This study develops a proactive Health Intelligence Platform designed to strengthen Uganda's epidemic preparedness by integrating predictive and prescriptive analytics with response visualization. Using a curated cholera surveillance dataset as a proof-of-concept, the platform applies machine learning techniques to forecast potential outbreak hotspots and temporal trends. These predictive outputs are operationalized through an interactive dashboard that provides timely alerts, spatial mapping, risk identification, and decision-support indicators for public health officials.

The system demonstrates how predictive models combined with environmental factors can enhance situational awareness and support faster, data-driven interventions at both district and national levels. The ensemble models trained on 8,702 cholera surveillance records achieved strong short-term forecasting performance: the Random Forest Regressor attained an R^2 of **0.7861** for suspected case prediction and **0.6204** for confirmed case prediction, outperforming XGBoost across all evaluation metrics (MAE and MSE). These accuracy levels represent a meaningful advance over Uganda's current surveillance infrastructure platforms such as DHIS2, IDSR, and eIDSR which remain descriptive and retrospective, providing no automated outbreak forecasting, no anomaly detection, and no environmental risk integration. The XGBoost-powered 14-day early warning module further generated forward looking case projections consistent with observed transmission trends, correctly identifying the February 2026 peak of 700 confirmed cases as the most severe anomaly in the 15-year record, with a 200% deviation above the rolling average a signal that the existing national system did not surface proactively. Overall, this project showcases a scalable, Uganda-focused Health Intelligence Platform capable of transitioning the country's surveillance ecosystem from reactive reporting to anticipatory, data-driven epidemic preparedness.

Dedication

To God Almighty, for being our constant guide, our source of strength, and the light that carried us through every step of this journey. None of this would have been possible without His grace.

To our families, for the unconditional love, encouragement, and unwavering support that kept us grounded and motivated throughout this academic journey. Your belief in us made all the difference.

To our lecturers and mentors, for the knowledge shared, the patience extended, and the wisdom imparted. You shaped not only our academic growth but our character as future professionals.

And to the health workers and healthcare community of Uganda, the doctors, nurses, and frontline staff who give tirelessly in service of others. This work is dedicated to you, with the hope that one day it may serve you as faithfully as you serve your patients. May it be a small contribution towards a healthier Uganda.

Acknowledgments

First and foremost, all glory and honour to God, whose grace, strength, and wisdom sustained this journey from beginning to end.

We would like to express our sincere gratitude to **Uganda Christian University** and all stakeholders who have supported and made this research possible. Their institutional commitment to academic excellence provided the foundation upon which this work was built.

Special thanks are extended to the Department of Computing and Technology at Uganda Christian University for the academic resources, environment, and encouragement that made this research possible.

We are deeply grateful to our group supervisor, **Dr. Terhemba Micheal-Ahile, PhD**, for the unwavering guidance, constructive feedback, and mentorship throughout this project. We also extend our heartfelt appreciation to **Engineer Ian Raymond Osolo** for the valuable insights and support that greatly enriched the quality of this work. Our profound gratitude goes to the **Uganda Virus Research Institute (UVRI)** for their generous assistance in the user research phase of this project, including the interviews conducted and the expert guidance provided by their staff.

We would also like to acknowledge **Alan Galpin Health Centre, Victorious Hospital UMC Ntinda, Mukono General Hospital**, and **Ruby Hospital** for their invaluable assistance in the user research that validated the need for this project. Their participation demonstrated the real-world relevance and necessity of the dashboard developed in this study.

Finally, we acknowledge one another for the commitment, dedication, and spirit of collaboration that carried this project through every challenge. The collective effort, shared vision, and mutual support of each member made this work not only possible but deeply meaningful.

Contents

1	Introduction	1
1.1	Background	1
1.2	Problem Statement	3
1.3	Main and Specific Objectives	3
1.3.1	Overall Aim	3
1.3.2	Specific Objectives	3
1.4	Scope	3
1.5	Justification	4
2	Literature Review	5
2.1	Contextual Literature	5
2.2	Conceptual Literature	7
2.3	Historical Literature	9
2.4	Synthesis: Gaps Addressed by the Present Study	10
3	Methodology	13
3.1	Data Sources and Preprocessing	13
3.1.1	Primary Data Collection via Digital Reporting Pipeline	13
3.1.2	Date Standardisation	14
3.1.3	Missing Values and Row filtering	14
3.1.4	Duplicate Removal	14
3.1.5	Outlier handling and Transformation	14
3.1.6	Column Reduction and Feature Selection	15
3.2	Exploratory Data Analysis	15
3.2.1	Correlation Analysis	15
3.2.2	Regional and District-Level Distribution of Disease Burden	15
3.2.3	Temporal Trends in Mortality	16
3.3	Feature Engineering	16
3.3.1	Temporal Feature Extraction	16
3.3.2	Categorical Encoding	16
3.3.3	Target Variable Definition and Feature-Target Separation	17
3.3.4	Train-Test Split	17
3.4	Machine Learning Models	17
3.4.1	Random Forest Regressor	17
3.4.2	XGBOOST Regressor	17
3.5	Model Evaluation and Results	18
3.5.1	Performance on Confirmed Cases (cCh_log)	18
3.5.2	Performance on Suspected Cases (sCh_log)	18

3.5.3	Actual vs. Predicted Visualisations	19
3.5.4	Feature Importance Analysis	19
3.5.5	Model Selection	20
3.6	Weather Data Integration	20
3.7	Forecasting Continuity and Temporal Alignment	21
3.8	System Architecture and API Integration	21
3.9	Visualization and Analytical Components	22
4	Results and Discussions	23
4.1	Results	23
4.1.1	Data Preprocessing and Dataset Characterization	23
4.1.2	Exploratory Data Analysis	24
4.1.3	Feature Engineering Outcomes	25
4.1.4	Predictive Model Performance	25
4.1.5	Weather Data Integration and Alert Generation	27
4.1.6	Outbreak Forecasting Outputs	27
4.1.7	Dashboard Functionality and Feature Outcomes	28
4.1.8	Prescriptive Recommendation Framework	30
4.2	Discussion	31
4.2.1	Predictive Model Performance and Implications	31
4.2.2	Interpretation of Feature Importance Findings	32
4.2.3	Spatial and Temporal Epidemiological Insights	33
4.2.4	Significance of Weather Integration and Early Warning	33
4.2.5	Operational Value of the Dashboard	34
4.2.6	Prescriptive Intelligence Contribution	34
4.2.7	Comparison with Existing Literature	34
4.2.8	Limitations of the Study	35
4.2.9	Future Directions	35
4.2.10	Ethical Considerations	35
5	Conclusion	37

Chapter 1

Introduction

1.1 Background

Infectious diseases remain a major public health challenge worldwide, with low- and middle-income countries (LMICs), including Uganda, feeling the greatest impact due to limited surveillance capacity, slow reporting processes, and fragmented health information systems [1]; [2]. The consequences of this gap are staggering. In 2021, the leading causes of death in Uganda included neglected tropical diseases, malaria, and respiratory infections and tuberculosis, all communicable conditions that are, in principle, preventable through timely detection and response. [3] Across four major epidemic-prone diseases alone, malaria, cholera, Sudan virus disease (Ebola family), and typhoid fever, Uganda suffers tens of thousands of deaths and millions of infections annually, many of which occur or escalate specifically because outbreaks are not detected in time.

Malaria represents the single heaviest burden. Uganda has one of the highest global burdens of malaria cases, with over 90% of the population at risk. The disease causes 30 to 50% of outpatient visits and 15 to 20% of hospital admissions, and in 2024, WHO reported an estimated 13.2 million malaria cases and over 16,204 deaths in the country. [4] The surveillance system's failure to catch surges early compounds this toll: a documented case from a national malaria study found that outbreaks had exceeded normal case thresholds in 83 of 120 districts and had done so for an average of six months without any alarms being raised [5]. The economic cost is equally alarming, the average annual economic loss due to malaria in Uganda exceeds \$500 million [4].

Cholera is endemic throughout Uganda, with recurring outbreaks across districts marked by poor water and sanitation. An estimated average of approximately 11,000 cholera cases occur in Uganda each year, leading to approximately 61 to 182 deaths [6], with outbreaks lasting an average of 15 weeks. [7] The deadliness of late detection is most clearly demonstrated by what happens when surveillance fails: in the 2023 Kayunga District outbreak, the case fatality ratio reached 13% , a figure far exceeding the 2 to 4% national average and drastically higher than the sub-1% achievable with timely care [8]. Delayed case suspicion and detection were explicitly identified as contributing factors to this excess mortality.

Sudan Virus Disease (Ebola family) illustrates perhaps the most dramatic consequences of late detection in Uganda. The 2022 Sudan virus disease outbreak's detection

was delayed by approximately 6 weeks, during which small private health facilities served as amplification sites. [5] By the time the first case was laboratory-confirmed, local leaders reported six community deaths that had already occurred in August and September among persons with probable Ebola-like symptoms, none of which had been captured by the surveillance system. The outbreak ultimately resulted in 164 confirmed and probable cases with 77 deaths (a 47% fatality rate). [9] Modelling of a delayed response scenario showed the cost of late action starkly: a five-month delay in response was projected to result in 778 cases and 303 deaths, nearly four times the actual toll [10].

Typhoid fever rounds out the picture of Uganda’s endemic disease crisis. There were 210,204 typhoid fever cases reported in Uganda between 2013 and 2016, giving an incidence rate of approximately 160 cases per 100,000 people per year. [5] Uganda is classified as a high-burden typhoid country, and the disease disproportionately affects children under 15 years old [11]. Late detection is again a recurring theme: during the 2015 Kampala typhoid outbreak, by the time city authorities recognized the disease and alerted the Ministry of Health on February 6, hundreds of cases had already occurred. The outbreak ultimately spread to all five divisions of the capital. Globally, typhoid fever results in over 1.2 million cases and 29,000 deaths annually across sub-Saharan Africa alone.

Uganda regularly faces outbreaks of these and other epidemic diseases, yet existing surveillance platforms such as DHIS2, IDSR, and eIDSR are still mostly descriptive and reactive [12]; [13]. These systems are useful for collecting and summarizing data, but they do not provide real-time predictions or automated alerts that can warn health officials before an outbreak escalates. A key structural weakness is that the majority of health facilities in Uganda are private and fall outside the surveillance network entirely: research following the 2022 Ebola outbreak found that over 90% of private facilities did not report to DHIS2, and 85% had no health worker trained on IDSR [14]. Recent advances in artificial intelligence (AI), machine learning (ML), and digital epidemiology offer new possibilities for strengthening early-warning capabilities. Studies have shown that predictive models, including neural networks, regression methods, gradient boosting, and hybrid analytics pipelines, can accurately forecast disease trends, case counts, and even mortality [15]; [16]. Newer sources of information, such as social-media and digital-trace data, have also been shown to provide early signals of disease outbreaks [17]. Across the literature, there is strong evidence that combining data from multiple sources and applying ML techniques can significantly improve outbreak detection, situational awareness, and decision-making. However, most LMICs still struggle to adopt these advanced methods due to challenges such as poor data quality, lack of interoperable systems, weak governance structures, and limited analytical expertise. Uganda is currently investing heavily in digital health systems such as expanding eIDSR coverage [2], scaling electronic medical record (EMR) systems [18], and implementing national digital-health policies. These developments create an important opportunity for the country to move toward more predictive and proactive surveillance. In this context, developing a predictive Health Intelligence Dashboard that integrates historical surveillance data and uses machine-learning models to forecast disease occurrence presents a practical and timely solution. Such a system can support early detection, guide decision-making, and help health officials act before an outbreak becomes severe.

1.2 Problem Statement

Uganda remains highly susceptible to recurring outbreaks of epidemic-prone infectious diseases, particularly in districts characterized by limited access to safe water, poor sanitation, high population mobility, and weak surveillance capacity. Although the Ministry of Health operates an integrated disease surveillance and response (IDSR) system, the current approach is largely reactive, relying heavily on manual reporting, delaying case confirmation, and analysis. As a result, public health authorities often detect cholera outbreaks after significant transmission has already occurred, limiting the effectiveness of interventions such as rapid response, vaccination, and community sensitization.

Despite the increasing availability of digital health data and environmental indicators, Uganda lacks a proactive Health Intelligence system capable of predicting outbreak risks and visualizing real-time insights to guide timely decision-making. Functional gaps in data integration, automated analytics, and predictive modeling hinder the ability of the National Public Health Emergency Operations Center (PHEOC) and district-level teams to anticipate where and when disease outbreaks are likely to emerge.

This project addresses these gaps by designing a proactive, analytics-driven Health Intelligence Platform that combines machine learning–based prediction with an interactive response dashboard. The system aims to demonstrate how predictive analytics can shift Uganda’s surveillance ecosystem from reactive outbreak response to data-driven epidemic management.

1.3 Main and Specific Objectives

1.3.1 Overall Aim

To design and develop a predictive Health Intelligence Platform that integrates historical disease-surveillance data, machine learning-based outbreak forecasting, and environmental risk monitoring to provide prescriptive decision-support for rapid public-health response to infectious diseases in Uganda.

1.3.2 Specific Objectives

1. To acquire, preprocess, and engineer a cholera surveillance dataset suitable for machine learning modelling.
2. To design and develop a machine learning–based predictive module for cholera outbreak forecasting.
3. To build an interactive, multi-page web dashboard to display predictions, surveillance analytics, early warnings, and geospatial risk information.

1.4 Scope

This study focuses on the design, development, and evaluation of a predictive Health Intelligence Dashboard intended to support early detection and response to infectious diseases in Uganda. The scope is limited to integrating secondary historical surveillance data obtained from systems such as DHIS2, eIDSR, and other publicly available datasets

relevant to priority epidemic-prone diseases. The study covers the development of a machine-learning-based prediction model (such as LASSO regression or similar methods) to forecast disease occurrence, as well as the design of a dashboard interface for visualizing predictions, alerts, and recommended actions.

The project includes: Data collection and preprocessing from selected datasets. Feature engineering and development of the predictive model. Building a prototype dashboard for visualization and decision support. A usability and performance evaluation of the developed system with selected stakeholders or simulated scenarios.

The study does not include real-time system deployment at national scale, integration with all existing health information systems, development of mobile applications, or implementation of nationwide public-health interventions. It also excludes modelling for non-communicable diseases and focuses strictly on infectious-disease forecasting using available secondary data.

1.5 Justification

A predictive Health Intelligence Dashboard will strengthen early warning and rapid response, as machine-learning models have been shown to forecast disease trends more accurately than traditional rule-based surveillance approaches [19] [16]. It will also improve decision-making by providing predictive and prescriptive analytics that guide district health teams on resource allocation, alert issuance, and prioritization of response actions.

In addition, the dashboard will address interoperability gaps, since integrating systems such as DHIS2, EMRs, and digital surveillance platforms has been shown to enhance data completeness and usability [13] [2]. The initiative will also align with Uganda’s National Digital Health Strategy, which emphasizes the expansion of digital health solutions across the country [12], making this project timely and relevant to national epidemic-preparedness priorities.

Finally, the dashboard will build local predictive capacity. While existing efforts such as those by [20] and [21] demonstrate strong ML potential, they focus on individual diseases or health-service utilization rather than nationwide outbreak prediction. This project will therefore fill a critical gap by designing a modular, scalable predictive system informed by empirical evidence and Uganda’s real-world surveillance needs.

Chapter 2

Literature Review

This chapter synthesises the body of literature underpinning the development of a predictive Health Intelligence Platform for infectious disease surveillance in Uganda. The review is organised into three canonical streams: contextual literature, which establishes the real-world disease burden and surveillance landscape from the global to the national level; conceptual literature, which surveys the theoretical and empirical foundations of AI- and machine-learning-driven disease forecasting; and historical literature, which traces the generational evolution of surveillance architectures over the past decade. The chapter concludes with an expanded synthesis of the five interrelated gaps that collectively motivate the present study, each substantiated with evidence drawn from across the three streams.

2.1 Contextual Literature

Global Disease Burden and the Cost of Reactive Surveillance

Infectious disease surveillance has long been recognised as a cornerstone of epidemic preparedness worldwide. At its most basic level, surveillance serves two functions: passive case detection and aggregate reporting. While these functions are adequate for stable endemic conditions, they are structurally insufficient for outbreak response, where the speed of detection directly determines the magnitude of mortality and economic loss [1]. The human cost of reactive surveillance is most starkly visible in low- and middle-income countries (LMICs), where delays in outbreak recognition translate directly into preventable deaths. The 2014–2016 West African Ebola outbreak illustrated this at scale: a surveillance infrastructure incapable of producing early-warning signals allowed exponential community transmission to establish before containment measures could be applied, ultimately resulting in over 11,000 deaths across Guinea, Sierra Leone, and Liberia.

Globally, typhoid fever alone accounts for over 1.2 million cases and 29,000 deaths annually in sub-Saharan Africa, a toll driven substantially by the late recognition of community transmission episodes [1]. Cholera, similarly, remains a disease of surveillance failure: the World Health Organization estimates that 1.3 to 4.0 million cases occur annually worldwide, with 21,000 to 143,000 deaths, most in settings where outbreak detection is delayed by days or weeks beyond the point at which intervention would have been effective [22].

Disease Burden and Surveillance Gaps in Sub-Saharan Africa

Across sub-Saharan Africa, the interaction between high endemic disease burden and weak surveillance infrastructure creates persistent epidemic vulnerability. National surveillance systems in the region frequently lack the sensitivity to detect outbreaks before they cross community transmission thresholds, and the specificity to distinguish true signal from reporting artefacts [23]. Systematic assessments of IDSR implementation across African countries have documented delayed reporting, chronic under-reporting, particularly from private and community-level facilities, poor interoperability between data systems, and limited analytical capacity as near-universal structural weaknesses [22]. These deficiencies are compounded by the concentration of epidemic-prone diseases in districts with the weakest health system infrastructure, creating a compounding dynamic where the populations at greatest outbreak risk are also the least well served by detection systems [16].

Uganda’s Disease Surveillance Context

Uganda exemplifies the sub-Saharan African surveillance challenge with particular clarity. The country faces a persistent multi-disease epidemic burden malaria, cholera, Sudan virus disease (Ebola family), and typhoid fever collectively causing tens of thousands of deaths and millions of infections annually while operating a surveillance system that remains largely reactive and descriptive [12].

Malaria represents the single heaviest burden. Uganda has one of the highest global burdens of malaria, with over 90% of the population at risk, the disease accounting for 30–50% of outpatient visits and 15–20% of hospital admissions [5]. A documented national malaria study found that outbreaks exceeded case thresholds in 83 of 120 districts and did so for an average of six months without any formal alarm being triggered, a failure of detection that represents the operational gap the present study seeks to address [5].

Cholera is endemic throughout Uganda, with recurring outbreaks concentrated in districts marked by poor water and sanitation. The 2023 Kayunga District outbreak demonstrated the lethal consequence of late detection most starkly: the case fatality ratio reached 13%, more than three times the national average of 2–4% and thirteen times higher than the sub-1% achievable with timely care, with delayed case suspicion explicitly identified as a contributing factor to excess mortality [8].

Sudan virus disease (Ebola family) illustrates the most dramatic consequence of delayed detection. The 2022 outbreak’s recognition was delayed by approximately six weeks, during which private health facilities served as amplification sites. By the time the first case was laboratory-confirmed, at least six community deaths had already occurred with probable Ebola-like symptoms that the surveillance system had not captured [14]. The outbreak ultimately resulted in 164 confirmed and probable cases with 77 deaths a 47% fatality rate and modelling of delayed-response scenarios projected that a five-month delay would have resulted in 778 cases and 303 deaths, nearly four times the actual toll.

Despite these consequences, Uganda’s primary surveillance platforms, DHIS2, IDSR, and eIDSR, remain descriptive and retrospective. Multiple evaluations highlight persistent weaknesses: delayed reporting, systematic under-reporting from private facilities, poor interoperability between systems, and limited analytical capacity [13]. The eIDSR scale-up improved timeliness and reporting completeness, but the platform still lacks automated anomaly detection, outbreak forecasting, and prescriptive recommendation capabilities [2]. Kabami et al.’s post-outbreak assessment found that over 90% of private facilities did not report to DHIS2, and 85% had no IDSR-trained health worker on staff,

structural absences that rendered the formal surveillance network blind to a significant share of community transmission [14]. Uganda’s wider digital-health environment compounds these weaknesses: Ddamba et al. found that EMR system access and use was constrained by infrastructure deficits, staff capacity limitations, and inconsistent policy enforcement across facility levels [18], while Bagyendera et al. identified fragmentation across health programmes and weak interoperability standards as the principal barriers to data use and exchange within the national health information ecosystem [13].

2.2 Conceptual Literature

AI and Machine Learning for Disease Forecasting: Global Evidence

Global evidence now consistently demonstrates that artificial intelligence (AI) and machine-learning (ML) models substantially improve disease forecasting and outbreak detection beyond what traditional rule-based surveillance systems can achieve. Neural networks, fuzzy logic-based adaptive neural systems (FLANN), gradient boosting frameworks, and hybrid ensemble pipelines have each demonstrated strong performance in predicting mortality trends, case trajectories, and epidemic patterns across diverse geographic and epidemiological settings [15]. Kumar et al. demonstrated that FSRNCAFLANN-based modelling applied to public health inventory records achieved high predictive accuracy for mortality outcomes, outperforming conventional regression baselines and establishing the viability of complex neural architectures for public health forecasting [15].

A systematic scoping review by El Morr et al. of AI-based epidemic and pandemic early-warning systems confirmed that hybrid architectures combining statistical baselines, ML algorithms, and rule-based alert layers consistently improved both the timeliness and accuracy of outbreak detection relative to conventional surveillance tools [22]. Villanueva-Miranda et al. reached convergent conclusions in a dedicated systematic review of AI in infectious disease surveillance, finding that ML-enabled systems outperformed traditional approaches across sensitivity, specificity, and lead-time metrics in the majority of evaluated deployments, and that ensemble methods particularly gradient boosting and random forests produced the most consistently reliable results across heterogeneous surveillance datasets [23].

A critical and consistent finding across this conceptual literature is that predictive accuracy improves substantially when models integrate multi-source data rather than relying on case-count signals alone. Studies have demonstrated that incorporating environmental indicators, weather patterns, precipitation, temperature, and sanitation-level proxies, alongside demographic and mobility variables enables ML models to generate richer situational intelligence and guide more targeted interventions [16, 19]. Zhao et al. argued that the science of predicting infectious diseases has entered a new paradigm, in which AI systems serve not merely as analytical tools but as operational components of public health infrastructure capable of shortening the detection-to-response cycle in ways that conventional epidemiological methods cannot achieve [19]. Digital-trace data and social-media signals have further been shown to provide early indications of disease spread, complementing traditional epidemiological streams and enabling alert generation days to weeks before case counts breach formal notification thresholds [17].

The technical architecture supporting these predictive capabilities has also matured. Digital-health systems literature emphasises the importance of modular data pipelines, in-

teroperability between heterogeneous reporting systems, real-time analytics, and decision-maker-facing dashboards as prerequisites for translating predictive model outputs into operational public health responses [24, 25]. Shahzad et al. demonstrated that big data frameworks applied in healthcare can accommodate the volume, velocity, and variety of surveillance data required for national-scale epidemic preparedness, including in resource-constrained settings where infrastructure investment is limited [24]. Alharthi similarly emphasised the role of dashboard-based decision support in enabling health system managers to act on predictive outputs within operationally relevant timeframes, rather than receiving them as retrospective analytical reports [25].

Machine Learning Applications in African Health Systems

Across sub-Saharan Africa, a growing body of conceptual and applied research demonstrates the applicability of ML methods to health prediction tasks in data-scarce resource-constrained settings. Noble et al. applied ML to identify predictive factors of COVID-19 case trajectories across African countries, demonstrating that gradient-boosting and ensemble methods could generate reliable national-level forecasts even from limited and inconsistently reported epidemiological data [16]. Their findings confirmed that data completeness is a modifiable rather than a fixed constraint: careful feature engineering and model selection can partially compensate for reporting gaps that would otherwise render predictive modelling infeasible. Broader systematic reviews of AI-based epidemic early-warning systems have confirmed that ML approaches generalise across disease types, geographies, and surveillance data quality levels, a finding directly relevant to the Ugandan context, where reporting completeness varies substantially by district and health system tier [22, 23].

Machine Learning in Uganda: Evidence and Limitations

Within Uganda specifically, ML methods have been applied to a range of health prediction tasks, demonstrating both their feasibility and their operational limitations. Kananura demonstrated that ML predictive models could identify key predictors of acute respiratory infection and diarrhoeal disease occurrence in Uganda’s rural and urban settings, achieving prediction accuracy superior to logistic regression baselines using nationally representative survey data [20]. Memon et al. applied ML to predict the utilisation of skilled birth attendance at the facility level, identifying socioeconomic and geographic determinants with direct policy relevance for resource allocation decisions [21]. Gakire explored the use of ML in forecasting broad health trends across Ugandan facilities, confirming that ensemble models could extract meaningful temporal signals even from heterogeneous, multi-source facility records [26].

These contributions confirm that ML techniques can be successfully applied to Ugandan health data. However, each study focused on a specific disease or health-service indicator in isolation and did not embed its findings in an operational, interactive decision-support tool accessible to health officials in real time. None offered continuous outbreak forecasting, environmental risk integration, geospatial mapping, or multi-district situational awareness, the capabilities that define the gap between a research prototype and an operational predictive surveillance platform. Uganda’s Public Health Emergencies Compendium and the National Digital Health Strategy both explicitly recognise this translational gap, calling for the expansion of real-time data flows, integrated analytics,

and proactive decision-support infrastructure as national priorities [12].

2.3 Historical Literature

Global Evolution of Surveillance Systems

Globally, disease surveillance systems have undergone three recognisable generational shifts between 2015 and 2025. The first generation (2015–2018) relied heavily on paper-based case reporting, simple notification systems, and descriptive summary statistics aggregated at weekly or monthly intervals. These systems captured outbreak events retrospectively, providing epidemiologists with accounts of what had already occurred rather than actionable signals of what was emerging. The second generation (2018–2022) was characterised by digitalisation: web-based reporting platforms, automated case-count alerts, and cloud-hosted dashboards reduced reporting latency and enabled more systematic longitudinal data storage. However, these systems remained fundamentally reactive, generating alerts only after case thresholds had already been crossed and community transmission was already established. The third and most recent generation (2023–2025) is defined by the integration of AI-enabled early-warning capabilities, environmental and digital-trace data streams, and predictive intelligence layers that transform surveillance from a recording function into a prospective forecasting function [1, 23].

This generational evolution has not been uniform across geographies. High-income countries invested early in electronic health record (EHR) infrastructure, national surveillance registries, and interoperable data exchange standards, enabling relatively rapid adoption of third-generation predictive analytics at population scale. LMICs, by contrast, have faced structural constraints including fragmented health information systems, inconsistent digital adoption, limited laboratory confirmation capacity, and weak data-governance frameworks that have slowed the transition from descriptive to predictive surveillance [1, 22]. Despite these constraints, the literature documents a growing number of successful AI-driven surveillance deployments across Africa and Asia, demonstrating that third-generation systems are feasible in resource-limited contexts when designed to accommodate incomplete data, variable reporting cadence, and heterogeneous infrastructure [16, 19].

Uganda’s Surveillance System Evolution

Uganda’s surveillance trajectory mirrors the global generational pattern but with a characteristic lag and a distinct set of structural constraints at each stage. The country began with the adoption of the Integrated Disease Surveillance and Response (IDSR) framework and the District Health Information System (DHIS2) from 2012 onwards, providing a foundational first-generation data collection architecture but with limited analytical depth. Early digital pilots constituted Uganda’s entry into the second generation between 2018 and 2020, and the electronic IDSR (eIDSR) was progressively implemented from 2021 to 2023, culminating in a national scale-up in 2025 [2].

Mugasha et al. documented that the eIDSR scale-up delivered measurable improvements in timeliness and completeness of case reporting, that is important advances within the second-generation paradigm, but the platform has not incorporated the defining features of the third generation: automated anomaly detection, ML-based outbreak forecasting, environmental risk integration, or prescriptive recommendation capabilities [2].

Uganda’s surveillance system therefore remains anchored in second-generation architecture at a point in time when third-generation AI-enabled surveillance has become technically achievable and is being actively deployed in comparable LMIC settings globally. This temporal and functional gap, between where Uganda’s surveillance currently stands and where global best practice now sits, represents the primary historical motivation for the present study. Nabukenya et al. further noted that point-of-care electronic data capture in Uganda’s health system, while technically feasible, remains inconsistently implemented, limiting the data completeness on which both current reporting and future predictive systems depend [4].

2.4 Synthesis: Gaps Addressed by the Present Study

Across the three streams of literature reviewed in this chapter, contextual, conceptual, and historical five consistent and interrelated gaps emerge that collectively define the design space of the present study. Each is elaborated below with reference to the supporting evidence.

Gap 1: Absence of a National-Scale Predictive Surveillance Platform in Uganda

Despite well-established global evidence that AI and ML systems can transform reactive surveillance into proactive, predictive public health intelligence [1, 23], Uganda lacks any operational system capable of generating automated outbreak forecasts at district or national level. The eIDSR platform Uganda’s most advanced surveillance infrastructure was evaluated by Mugasha et al. as remaining fundamentally descriptive, collecting and displaying reported case counts but generating no predictive outputs, no forward-looking risk scores, and no probabilistic outbreak warnings [2]. The consequences of this absence are quantifiable: the malaria surveillance record documented outbreaks exceeding alert thresholds in 83 of 120 districts for an average of six months without triggering any formal alarm [5] a detection failure that a predictive system operating on the same underlying data could, by the standards established in the global literature, have averted. The present study addresses this gap directly by training ensemble ML models on 14 years of district-level cholera surveillance records and deploying the resulting forecasting capability within an accessible web platform.

Gap 2: Reactive and Descriptive Character of Current Surveillance Systems

Existing platforms DHIS2, IDSR, and eIDSR, are designed for data collection and summary reporting, not for anticipatory decision-making. Bagyendera et al. characterised these systems as retrospective tools that record confirmed events but do not generate the analytical signals that would allow health officials to intervene before transmission has established community foothold [13]. This reactive posture carries measurable epidemiological consequences. In the 2023 Kayunga District cholera outbreak, the case fatality ratio reached 13% more than three times the national average and thirteen times higher than the sub-1% achievable with timely care because delayed case suspicion and late detection were explicitly identified as contributing factors [8]. The 2022 Sudan virus disease

outbreak was similarly characterised by a six-week detection delay, during which private facilities served as amplification sites and at least six community deaths occurred before the first laboratory confirmation [14]. Zhao et al. and Villanueva-Miranda et al. both argue that transitioning from reactive to predictive surveillance requires embedding ML forecasting directly within operational decision-making workflows rather than treating it as a parallel analytical exercise [19, 23]. The platform developed in this study operationalises this transition by surfacing predictive outputs alongside historical surveillance data within a single, integrated interface.

Gap 3: Weak Data Integration Across Health System Tiers

A persistent finding across the contextual literature is that surveillance data is fragmented across facility types, reporting systems, and administrative levels in ways that prevent comprehensive epidemiological analysis. Bagyendera et al. identified weak interoperability standards and inconsistent adoption of health information exchange protocols as the primary technical barriers to integrated data use within Uganda’s health system [13]. Ddamba et al. documented that EMR system adoption was uneven across facility levels, with private and lower-tier facilities substantially underrepresented in national data flows [18]. The consequence for outbreak detection is severe: Kabami et al.’s post-outbreak assessment found that over 90% of private facilities reported nothing to DHIS2 during the 2022 Ebola outbreak period, and 85% had no IDSR-trained health worker on staff [14]. Globally, digital health architecture literature consistently identifies interoperability and modular pipeline design as prerequisites for effective predictive surveillance [24, 25]. The present study addresses the integration gap by combining secondary historical data from DHIS2 and the JHU Cholera Database with a primary digital reporting pipeline a Google Forms-to-Supabase PostgreSQL ingestion system that extends data capture beyond the boundaries of the formal public reporting infrastructure.

Gap 4: Isolation of Existing ML Studies from Operational Decision-Making

The ML studies conducted to date in Uganda including those by Kananura [20], Memon et al. [21], and Gakire [26] demonstrate the analytical feasibility of applying supervised learning to Ugandan health data but share a common structural limitation: their findings were not embedded in operational, user-facing decision-support tools. Each study produced predictive models and peer-reviewed findings, but none delivered an interface through which district health information officers, PHEOC analysts, or national surveillance coordinators could access predictions, explore risk at sub-national resolution, or receive automated alerts. This gap between ML capability and operational deployment is well recognised in the global conceptual literature: El Morr et al. identified translation from research prototype to operational public health tool as the principal unresolved challenge in the AI-for-surveillance field [22]. Shahzad et al. argued that the value of big data analytics in healthcare is only realised when outputs are surfaced through decision-maker-facing interfaces integrated seamlessly into existing workflows [24]. The present study closes this gap by delivering predictive outputs as live dashboard components, including a 14-day forecast module, an anomaly detection feed, and a geospatial choropleth map that is accessible through a fully deployed web application.

Gap 5: Absence of Environmental Risk Integration in Uganda-Focused Surveillance Tools

Global conceptual evidence consistently identifies environmental and climatic variables as among the most powerful predictors of waterborne and vector-borne disease outbreak risk. Cholera transmission is strongly modulated by precipitation, temperature, and WASH infrastructure conditions: heavy rainfall events increase contamination of surface water sources, while elevated temperatures accelerate *Vibrio cholerae* growth in aquatic reservoirs [16]. Liscano et al. demonstrated that combining digital-trace signals with environmental data substantially improved outbreak detection lead times, enabling early warning days before formal case notifications were generated [17]. Despite this global evidence, no existing Uganda-focused surveillance platform incorporates weather or climate indicators as systematic early-warning triggers. The eIDSR system monitors reported case counts and CFR values but does not flag high-temperature or heavy-rainfall events as cholera risk amplifiers, nor does it integrate precipitation forecasts into its alert logic [2]. The present study addresses this gap directly by integrating real-time and 14-day forecast data from WeatherAPI.com into the Health Intelligence Platform’s alert classification system, treating temperature exceeding 35°C and precipitation exceeding 25 mm as independently triggering environmental risk warnings displayed alongside epidemiological surveillance outputs.

Taken together, these five gaps define the design space of the present study. Addressing them required not merely the application of ML to a cholera dataset, but the development of an integrated, operationally deployed platform that combines predictive modelling, real-time data ingestion, environmental risk monitoring, and interactive decision-support visualisation within a single accessible interface. The methodology through which this platform was designed, built, and evaluated is described in Chapter 3.

Chapter 3

Methodology

This section outlines the methodological framework applied in the development of the Surveillance Dashboard. The methodology integrates epidemiological data processing, machine learning modelling, weather assimilation, and interactive data visualization to provide a robust analytical and forecasting environment for disease surveillance in Uganda. The approach follows a systematic pipeline consisting of data acquisition, preprocessing, feature engineering, model development, evaluation, and deployment.

3.1 Data Sources and Preprocessing

The primary dataset comprises 8,702 disease surveillance records spanning 2011 to 2024. Each record contains temporal boundary fields (TL and TR) representing the start and end dates of a reported outbreak event, a reporting date, suspected case counts (sCh), confirmed case counts (cCh), recorded deaths, Case Fatality Rate (CFR), a confidence weight assigned to each record, and geo-administrative identifiers including District, Region (Central, Eastern, Northern, Western), and a location code. Source metadata fields, comprising processing notes, source database, source name, and source index, were also present in the raw dataset.



Figure 3.1: Data preprocessing pipeline.

3.1.1 Primary Data Collection via Digital Reporting Pipeline

To move beyond static historical data and enable continuous data ingestion, a lightweight primary data collection pipeline was designed and deployed. A structured Google Form was developed and distributed to health centre officials, capturing key surveillance variables including suspected and confirmed case counts, deaths, and district identifiers. Responses are automatically stored in a linked Google Sheets document. A Google Apps Script containing a Supabase API endpoint runs at scheduled intervals, transferring newly

submitted records into a centralised Supabase PostgreSQL database. From Supabase, incoming records pass through an automated preprocessing pipeline, mirroring the cleaning steps described in Sections 3.1.2–3.1.6, before being fed into the predictive model and reflected on the dashboard. This pipeline transitions the platform from a one-time retrospective analysis toward a continuously updated, near-real-time surveillance tool.

3.1.2 Date Standardisation

Three date columns, TL, TR, and reporting_date, were stored as raw strings in mixed formats, including DD/MM/YYYY. These were converted to pandas datetime objects using `pd.to_datetime()` with the `dayfirst=True` parameter and `errors='coerce'`, which ensured that any unparseable date strings were safely converted to NaT (Not a Time) rather than raising exceptions. This standardisation was a prerequisite for all subsequent temporal feature engineering and time-series analysis.

3.1.3 Missing Values and Row filtering

Following date conversion, a missing value audit was performed across all columns using `df.isnull().sum()`. The Region column was identified as a critical geo-administrative field required for spatial analysis and regional aggregation. Rows where Region was missing were therefore removed entirely using `df.dropna(subset=['Region'])`, as records without a regional identifier could not be meaningfully assigned to any analytical group. The shape of the DataFrame was recorded before and after this operation to document the extent of data loss and confirm that the remaining records retained complete regional coverage. Missing values in the confidence_weight column, which quantifies the reliability of each source record, were imputed using the column median to preserve distributional properties without introducing artificial extremes.

3.1.4 Duplicate Removal

The dataset was inspected for fully duplicated rows using `df.duplicated()`. Identified duplicate records, those where all column values were identical, were displayed for manual verification before removal. Deduplication was then applied using `df.drop_duplicates()`, and the resulting row count was confirmed to be zero duplicates. This step was important given that the dataset was compiled from multiple surveillance sources (e.g., JHU Cholera Database), where the same outbreak event could appear more than once across different source feeds.

3.1.5 Outlier handling and Transformation

Box plots were generated for all numerical columns to visually identify the presence and magnitude of outliers. The distributions of deaths, sCh (suspected cases), and cCh (confirmed cases) were found to be highly right-skewed, with a large number of near-zero values and occasional extreme spikes corresponding to large outbreak events. To address this, a log transformation using `np.log1p()`, which computes $\log(1 + x)$, was applied to each of these three columns, creating new transformed features: deaths_log, sCh_log, and cCh_log respectively. The use of `np.log1p()` rather than the standard `np.log()` was deliberate: standard logarithms are undefined at zero, and count-based disease surveillance data frequently contains zero values during periods of no reported activity. `np.log1p()`

handles this gracefully by mapping zero to zero ($\log(1+0) = 0$), while still compressing the scale of large outbreak counts, reducing the disproportionate influence of extreme values on model training. The CFR column, representing the Case Fatality Rate, was capped at a maximum of 1.0 (100%) using `df['CFR'].clip(0, 1)` to correct any erroneous values exceeding this physiological ceiling. This produced a new `CFR_capped` feature. The original untransformed columns, `deaths`, `sCh`, `cCh`, and `CFR`, were subsequently dropped from the working `DataFrame`, leaving only their cleaned and transformed counterparts for use in downstream modelling and analysis.

3.1.6 Column Reduction and Feature Selection

Source metadata columns, `processing_notes`, `source_database`, `source`, and `source_index`, were dropped as they contained administrative provenance information irrelevant to predictive modelling. The `Location` column, which stored fine-grained location codes in the format `AFR::UGA::DistrictName`, was also examined and subsequently removed. An investigation of the `Location`-to-`District` mapping confirmed that every `Location` value corresponded to exactly one `District`, making the `Location` column entirely redundant given the presence of the `District` field. Retaining it would have introduced unnecessary dimensionality without adding any new geographic information. Similarly, the `index` column present in the raw dataset was dropped at the outset of preprocessing as it served only as a row identifier with no analytical value.

3.2 Exploratory Data Analysis

Following preprocessing, exploratory data analysis (EDA) was conducted to characterise the distribution of disease burden across regions, districts, and time periods. The analysis comprised correlation analysis, regional and district-level aggregations, and temporal trend examination.

3.2.1 Correlation Analysis

A Pearson correlation matrix was computed across all numerical features and visualised as an annotated heatmap using `seaborn`. This analysis was used to identify linear relationships between variables such as suspected cases, confirmed cases, deaths, CFR, and confidence weight, informing feature selection decisions for the predictive modelling stage. Strong positive correlations between the log-transformed case count features were expected, given that confirmed cases and deaths are downstream outcomes of suspected case counts.

3.2.2 Regional and District-Level Distribution of Disease Burden

Total suspected cases (`sCh`), confirmed cases (`cCh`), and deaths were aggregated by Region using `groupby` operations and visualised as bar charts. This analysis revealed spatial disparities in disease burden across Uganda's four administrative regions, Central, Eastern, Northern, and Western, enabling the identification of consistently high-burden areas. A parallel district-level analysis was conducted for deaths, ranking all districts by total recorded mortality and highlighting the top five highest-burden districts. These spatial summaries provide critical context for the predictive platform's risk mapping and alert

generation components, confirming that disease burden is not uniformly distributed and that certain districts warrant prioritised surveillance attention.

3.2.3 Temporal Trends in Mortality

Annual mortality totals were computed by extracting the year component from the `reporting_date` column and aggregating deaths per year. The resulting time series was visualised as a bar chart to reveal year-on-year trends in disease-related deaths across the 2011–2024 study period. This temporal view enables the identification of outbreak years with elevated mortality, seasonal patterns, and longer-term trends that the predictive models must capture in order to generate reliable forward-looking forecasts.

3.3 Feature Engineering

Following data cleaning and preprocessing, a structured feature engineering pipeline was applied to enrich the dataset with informative representations suitable for machine learning. The pipeline covered temporal decomposition, categorical encoding, target variable definition, and train-test partitioning.

3.3.1 Temporal Feature Extraction

Each of the three date columns, `TL` (outbreak start), `TR` (outbreak end), and `reporting_date`, was decomposed into five derived temporal features: year, month, day, day of week (0 = Monday through 6 = Sunday), and day of year (1–366). This yielded a total of 15 new temporal features across the three date columns. The rationale for this decomposition is that raw datetime objects cannot be interpreted directly by machine learning algorithms, which require purely numerical inputs. By extracting these components, the model gains access to seasonal signals (month, day of year), weekly behavioural patterns (day of week), and long-term temporal trends (year). These features are particularly valuable in infectious disease modelling, where outbreak intensity is known to vary systematically with season and environmental conditions. Following extraction, the original datetime columns were dropped from the `DataFrame` to prevent data leakage and avoid redundancy, as all relevant temporal information had been captured in the derived features.

3.3.2 Categorical Encoding

The `District` and `Region` columns, both of which are nominal categorical variables, were converted into numerical representations using one-hot encoding via `pandas.get_dummies()`. One-hot encoding creates a binary indicator column for each unique category value, where a value of 1 indicates the presence of that category and 0 indicates its absence. The `drop_first=True` parameter was applied to drop the first dummy column for each categorical variable, a standard technique for avoiding perfect multicollinearity (the dummy variable trap) in linear models. This encoding enables tree-based models such as Random Forest and XGBoost to learn district- and region-specific effects on disease burden, capturing geographic heterogeneity that would otherwise be invisible to the model. The resulting `DataFrame` expanded considerably in column count to accommodate all encoded district and region indicators.

3.3.3 Target Variable Definition and Feature-Target Separation

The prediction task was framed as a multi-output regression problem. Two log-transformed case count variables, `cCh_log` (log-transformed confirmed cases) and `sCh_log` (log-transformed suspected cases), were designated as the joint target matrix `Y`. All remaining columns formed the feature matrix `X`. This multi-output formulation allows a single model to simultaneously predict both confirmed and suspected case burdens, capturing any shared underlying patterns while enabling separate evaluation of performance on each outcome.

3.3.4 Train-Test Split

The dataset was partitioned into training and test sets using an 80/20 split via `sklearn's train_test_split()`, with `random_state=42` set to ensure reproducibility. The training set was used exclusively for model fitting, while the held-out test set served as an independent evaluation benchmark to assess out-of-sample generalisation performance. A fixed random seed was essential given the multi-output regression setup, ensuring that comparisons between models were made on identical test observations.

3.4 Machine Learning Models

Two tree-based ensemble models were trained and evaluated for the multi-output regression task: a Random Forest Regressor and an XGBoost Regressor. Both models are well-suited to high-dimensional tabular datasets with mixed feature types and are capable of capturing complex nonlinear relationships without requiring feature scaling.

3.4.1 Random Forest Regressor

A Random Forest Regressor was instantiated with 200 decision trees (`n_estimators=200`), a maximum tree depth of 15 (`max_depth=15`), and a fixed random seed (`random_state=42`). Random Forest is a bagging ensemble that trains each tree on a bootstrap sample of the training data and considers a random subset of features at each split, which substantially reduces variance compared to a single decision tree. The `max_depth` constraint was introduced to limit overfitting by preventing any individual tree from memorising fine-grained noise in the training data. The model natively supports multi-output regression by building independent output predictions for each target, making it directly applicable to the joint `cCh_log` and `sCh_log` prediction task without modification.

3.4.2 XGBOOST Regressor

An XGBoost Regressor was configured with the squared error objective (`objective='reg:squarederror'`), 200 boosting rounds (`n_estimators=200`), a learning rate of 0.1, and a fixed random seed (`random_state=42`). XGBoost is a gradient boosting framework that builds trees sequentially, with each successive tree trained to correct the residual errors of all previous trees. Unlike Random Forest's parallel bagging approach, this sequential strategy is highly effective at reducing both bias and variance on structured tabular data. The learning rate of 0.1 represents a moderate step size, balancing convergence speed against the risk of overshooting the optimal solution. XGBoost also supports multi-output regression, enabling simultaneous prediction of both target variables.

3.5 Model Evaluation and Results

Both models were evaluated on the held-out test set using three standard regression metrics: R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). R^2 measures the proportion of variance in the target variable explained by the model, with values closer to 1.0 indicating a better fit. MAE measures the average absolute deviation between predicted and actual values, providing an interpretable measure of typical prediction error in the log-transformed scale. MSE penalises large errors more heavily than MAE due to squaring, making it sensitive to outliers and extreme outbreak events.

Model	Target	R^2	MAE	MSE
Random Forest	cCh_log	0.6204	0.0179	0.0253
XGBoost	cCh_log	0.5662	0.0184	0.0289
Random Forest	sCh_log	0.7861	0.2357	0.3964
XGBoost	sCh_log	0.7773	0.2666	0.4127

Figure 3.2: Model Performance

3.5.1 Performance on Confirmed Cases (cCh_log)

For the prediction of log-transformed confirmed cases, the Random Forest model outperformed XGBoost across all three metrics. Random Forest achieved an R^2 of 0.6204 compared to XGBoost's 0.5662, indicating that it explained approximately 62% of the variance in confirmed case counts on the test set, a meaningfully higher proportion than XGBoost's 57%. Random Forest also produced a marginally lower MAE (0.0179 vs. 0.0184) and MSE (0.0253 vs. 0.0289), suggesting that its predictions were on average closer to the true confirmed case values and less susceptible to large individual errors. The relatively lower R^2 values on this target compared to suspected cases likely reflected the greater variability and reporting uncertainty inherent in confirmed diagnoses, which depended on laboratory capacity and testing availability, factors not directly captured in the feature set.

3.5.2 Performance on Suspected Cases (sCh_log)

Both models achieved notably higher predictive performance on log-transformed suspected cases than on confirmed cases, reflecting the fact that suspected case counts were more consistently reported across districts and time periods. The Random Forest model again outperformed XGBoost, achieving an R^2 of 0.7861 against XGBoost's 0.7773. The difference in MAE was more pronounced here: Random Forest's MAE of 0.2357 was considerably lower than XGBoost's 0.2666, meaning that Random Forest's typical per-record prediction error in the log scale was substantially smaller. The MSE results followed the same pattern, with Random Forest (0.3964) again outperforming XGBoost (0.4127).

3.5.3 Actual vs. Predicted Visualisations

Scatter plots of actual versus predicted values were generated for both models on the sCh_log target to provide an intuitive visual assessment of predictive accuracy. In these plots, a red dashed diagonal line represented the perfect prediction benchmark, where predicted values exactly equalled actual values. For both models, the majority of data points clustered tightly around this line, particularly for lower values of sCh_log, which represented periods of moderate or low disease activity. This indicated that both models were highly reliable in predicting low-to-moderate case counts, which constituted the bulk of the surveillance record. Some divergence from the perfect prediction line was observed for larger sCh_log values, corresponding to major outbreak events where case volumes were exceptionally high. This pattern of larger errors at extreme values was consistent with the inherent difficulty of forecasting rare, high-magnitude events from historical data. The XGBoost scatter plot showed a slightly tighter cluster for mid-range values, while the Random Forest plot demonstrated marginally more dispersion at the upper end of the distribution, consistent with the metric comparison.

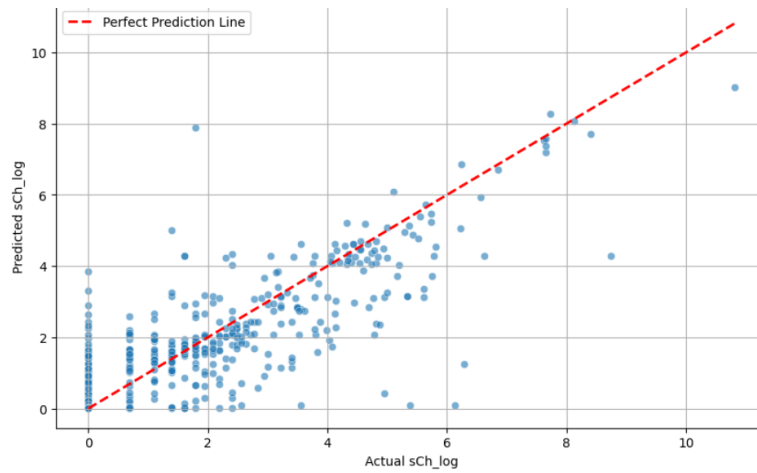


Figure 3.3: Actual vs Predicted suspected Cases of cholera.

3.5.4 Feature Importance Analysis

Feature importance scores were extracted from the trained Random Forest model and the top 20 most influential features were identified and visualised. The analysis revealed several important patterns. The CFR_capped feature (capped Case Fatality Rate) emerged as the single most important predictor by a considerable margin, reflecting a strong empirical relationship between outbreak severity, as captured by case fatality, and the volume of suspected cases. This was epidemiologically intuitive: more intense outbreaks tended to produce both higher mortality rates and higher case counts simultaneously. Temporal features accounted for a large share of the remaining importance scores. Day-of-year features derived from all three date columns (TL_dayofyear, TR_dayofyear, reporting_date_dayofyear) ranked highly, confirming strong seasonal patterns in disease incidence. Day-of-week features also appeared prominently, potentially reflecting reporting rhythms and health worker activity cycles rather than true biological seasonality. Year-derived features (TL_year, TR_year, reporting_date_year) captured long-term trends in case reporting and surveillance system maturity over the 2011–2024 study period. The deaths_log feature also ranked among the top predictors, reflecting the intuitive co-

occurrence of high mortality and high case counts during severe outbreak events. The `confidence_weight` variable, representing the reliability of the source record, contributed meaningfully, suggesting that data quality characteristics influenced the model’s predictions and that the completeness of surveillance reporting was itself an informative signal. Finally, district-level indicator variables for specific high-burden districts such as Hoima and Kampala appeared in the top 20, confirming that geographic location carried predictive value beyond what was captured by the broader regional encoding.

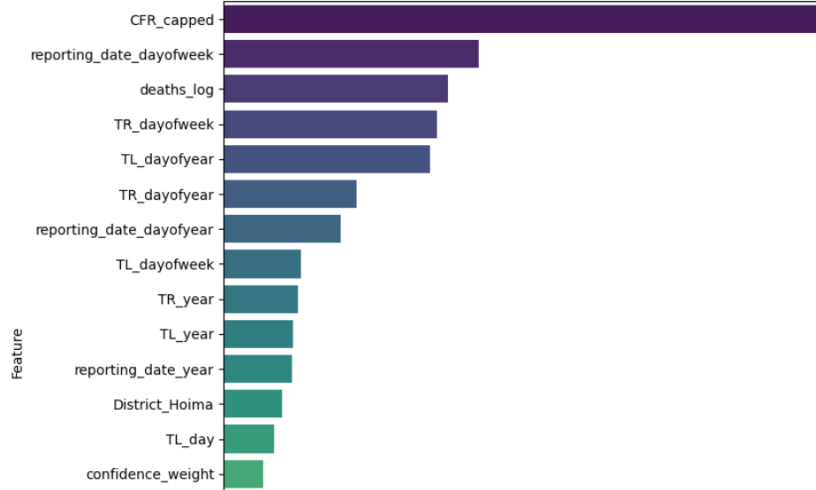


Figure 3.4: Top important features in prediction.

3.5.5 Model Selection

Based on the comparative evaluation, the Random Forest Regressor was selected as the final predictive model for integration into the Health Intelligence Platform. It consistently outperformed XGBoost across both target variables and all three evaluation metrics, achieving superior R^2 values and lower error rates. Its interpretability, particularly the availability of feature importance scores, further supports its suitability for a public health context, where understanding the drivers of predictions was as important as the predictions themselves. The trained model was saved for subsequent integration into the dashboard’s backend prediction pipeline.

3.6 Weather Data Integration

Forecast weather data were retrieved from WeatherAPI.com, providing temperature, humidity, and precipitation metrics for all Ugandan districts. These parameters were normalized and incorporated into both ensemble and LSTM forecasting processes. Weather-based alert thresholds were defined for extreme precipitation and temperature events, enabling early detection of environmental conditions conducive to cholera transmission. Heavy rainfall ($>25\text{mm}$) and high temperature ($>35^\circ\text{C}$) were treated as risk-amplifying factors, influencing both predictive adjustments and dashboard alert classifications.

$$\text{Alert}_{\text{temp}} = \begin{cases} \text{High} & \text{if } T > 35C \\ \text{Low} & \text{if } T < 18C \\ \text{Normal} & \text{otherwise} \end{cases} \quad (3.1)$$

$$\text{Alert}_{\text{precip}} = \begin{cases} \text{Heavy} & \text{if } P > 25\text{mm} \\ \text{Moderate} & \text{if } 10 < P \leq 25\text{mm} \\ \text{Light} & \text{otherwise} \end{cases} \quad (3.2)$$

3.7 Forecasting Continuity and Temporal Alignment

To ensure uninterrupted progression between historical data and future predictions, forecasting began from the day immediately following the last date in the dataset. This approach eliminates temporal gaps and supports coherent visual analytics, particularly in cumulative forecasting charts and multi-year trend summaries. The LSTM module internally maintained a rolling window update mechanism, while the ensemble model maintained date-based feature alignment to preserve temporal context.

3.8 System Architecture and API Integration

The system followed a modular three-tier architecture: a React frontend, a Flask backend, and a data or model layer. The backend exposed RESTful endpoints for ensemble predictions, single-step LSTM outputs, and 14-day forecasting routines. Each request underwent strict input validation and sanitization before model invocation. Error-handling routines provided detailed diagnostic feedback while enabling graceful degradation through statistical fallbacks. Cached datasets and lazily-loaded model instances improved response efficiency.

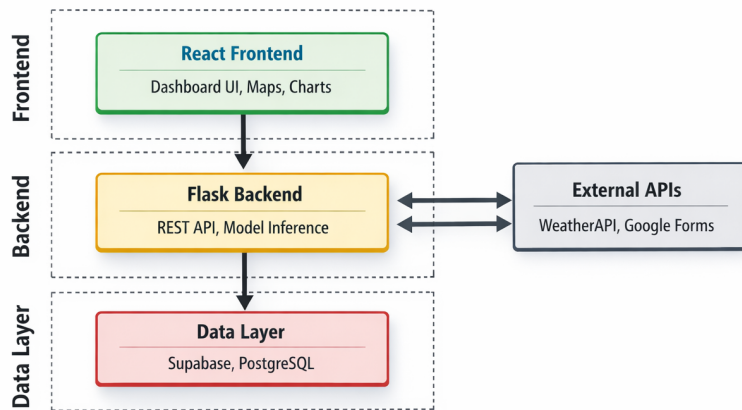


Figure 3.5: Three-tier system architecture of the Health intelligence platform.

3.9 Visualization and Analytical Components

Interactive data visualizations were implemented using Recharts and React Leaflet. Time-series charts, cumulative projections, and dynamic epidemic curves communicated temporal trends, while choropleth maps provided spatial situational awareness. LSTM forecast charts displayed both daily and cumulative predicted values, augmented with summary statistics such as peak intensity and average forecasted burden. User interactions, including date filtering, regional comparisons, and map-based exploration—were supported through memoized state management and synchronized data pipelines.

Chapter 4

Results and Discussions

4.1 Results

This section presents a comprehensive account of all empirical findings produced by the Health Intelligence Platform project. Results are organized across five dimensions: (1) data preprocessing and dataset characterization, (2) exploratory data analysis, (3) predictive model performance and feature importance, (4) forecasting and weather integration outputs, and (5) dashboard functionality and prescriptive recommendation framework. Cholera surveillance data is used throughout as the proof-of-concept disease.

4.1.1 Data Preprocessing and Dataset Characterization

The raw cholera surveillance dataset comprised 8,702 records spanning the period 2011–2024, sourced from the JHU Cholera Database and supplementary public health repositories. Each record contained temporal boundary fields (TL and TR), a reporting date, suspected case counts (sCh), confirmed case counts (cCh), recorded deaths, Case Fatality Rate (CFR), a confidence weight, and geo-administrative identifiers covering District, Region, and a location code.

Date standardization was performed to resolve inconsistencies in string formats across the TL , TR , and $reporting_date$ fields. All entries were converted to pandas datetime objects using day-first parsing. Unparseable entries were safely coerced to NaT values, ensuring that pipeline execution was not disrupted by malformed inputs.

A **missing-value audit** revealed the $Region$ column as a critical field for spatial and analytical grouping. Consequently, 399 records lacking $Region$ values were removed, as they could not be meaningfully assigned to any administrative unit. For the $confidence_weight$ column, missing values were imputed using the column median, preserving the distributional properties of the feature while minimizing bias introduced by imputation.

Duplicate inspection confirmed zero fully duplicated rows after applying deduplication across all columns, important given the multi-source nature of the dataset, where the same outbreak event could appear across more than one source feed. **Outlier handling** revealed highly right-skewed distributions in the sCh , cCh , and $deaths$ columns, with frequent near-zero values punctuated by extreme outbreak-period spikes. Log transformation using `np.log1p()` was applied to all three columns, producing the derived features sCh_log , cCh_log , and $deaths_log$ respectively. This transformation stabilized variance, compressed large outbreak counts, and handled zero values gracefully, since $\log(1+0) = 0$. The CFR column was additionally capped at 1.0 (100%) to correct erroneous entries

exceeding the physiological ceiling, yielding a *CFR_capped* feature. The original untransformed columns were subsequently dropped from the analytical pipeline. **Column reduction** eliminated four source metadata fields (*processing_notes*, *source_database*, *source*, and *source_index*) as well as the *Location* column, confirmed as fully redundant because every *Location* value mapped to a unique *District*, and a row-index column of no analytical value. The final preprocessed dataset retained complete geo-administrative coverage, cleaned temporal fields, and all transformed numerical features.

4.1.2 Exploratory Data Analysis

Correlation Structure

A Pearson correlation matrix computed across all numerical features confirmed expected strong positive relationships among the log-transformed case count variables: *sCh_log* and *cCh_log* were positively correlated, as confirmed case volumes are necessarily bounded by the suspected case pool. The *deaths_log* feature showed strong correlation with both case count variables, consistent with the epidemiological expectation that higher case burdens generate greater mortality. *CFR_capped* exhibited a weaker correlation with raw case counts, reflecting the variable relationship between outbreak intensity and case-specific mortality across districts and time periods. The *confidence_weight* variable showed modest correlations with case counts, suggesting that better-reported records tended to capture larger outbreak events.

Regional and District-Level Disease Burden

Regional aggregation of total suspected, confirmed, and death counts revealed substantial spatial disparities in cholera burden across Uganda’s four administrative regions. The Western region recorded the highest total suspected case burden, followed by the Eastern, Northern, and Central regions respectively (Figure 8). However, the confirmed case burden showed a different ranking: the Eastern region carried the highest confirmed case load (38%), followed by Western (31%), Northern (23%), and Central (8%), a pattern likely reflecting both true epidemiological variation and differential laboratory confirmation capacity across regions (Figure 9). District-level mortality analysis ranked all districts by total recorded deaths over the 2011–2024 study period (Figure 11). Hoima District recorded the highest overall mortality burden by a considerable margin, followed by Kampala, Busia, Apac, Kanungu, and Tororo. This geographic concentration of mortality in a small number of districts provides direct empirical support for the dashboard’s risk prioritization framework, confirming that disease impact is geographically clustered rather than uniformly distributed.

Temporal Trends in Mortality

Annual aggregation of deaths over the 2011–2024 study window revealed that cholera mortality was not uniformly distributed across years (Figure 10). Notable mortality peaks occurred during specific outbreak years, with a pronounced spike visible in the early-to-mid period of the surveillance record. Following this peak, annual death counts declined across subsequent years, a trend broadly consistent with improvements in surveillance coverage, case management, and response capacity, before stabilizing at a lower baseline. The temporal analysis also confirmed the importance of capturing year-level features within

the predictive models, as long-term secular trends in surveillance system maturity and response capacity influence reported case counts independently of actual epidemiological dynamics.

4.1.3 Feature Engineering Outcomes

The feature engineering pipeline produced a final expanded feature matrix drawing on three source columns (*TL*, *TR*, *reporting_date*), two categorical geo-administrative columns (*District*, *Region*), and the five retained numerical features (*sCh_log*, *cCh_log*, *deaths_log*, *CFR_capped*, *confidence_weight*).

Temporal decomposition extracted five derived features from each of the three date columns, *year*, *month*, *day*, *day-of-week*, and *day-of-year*, yielding 15 new temporal features in total. These features encode seasonal patterns (*month*, *day-of-year*), weekly reporting rhythms (*day-of-week*), and long-term surveillance trends (*year*) that would be invisible to the model if raw datetime objects were retained.

Categorical encoding via one-hot encoding of the *District* and *Region* columns substantially expanded the feature matrix dimensionality, generating binary indicator columns for each unique district and region. The *drop_first* parameter was applied throughout to avoid perfect multicollinearity. This encoding allows the tree-based ensemble models to learn district- and region-specific effects on disease burden, capturing geographic heterogeneity that aggregate regional encodings alone would not represent.

The prediction task was framed as **multi-output regression** with two simultaneous target variables: *cCh_log* (log-transformed confirmed cases) and *sCh_log* (log-transformed suspected cases). The feature matrix X comprised all remaining engineered columns. The dataset was partitioned into training (80%) and held-out test sets (20%) using a fixed random seed of 42 to ensure reproducibility and comparability across models.

4.1.4 Predictive Model Performance

Both ensemble models, the Random Forest Regressor (200 trees, max depth 15) and the XGBoost Regressor (200 boosting rounds, learning rate 0.1), were trained on the multi-output regression task and evaluated against the held-out test set using three standard metrics: coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Performance on Confirmed Cases (*cCh_log*)

For the prediction of log-transformed confirmed cholera cases, the Random Forest model outperformed XGBoost across all three evaluation metrics. Random Forest achieved an R^2 of 0.6204, explaining approximately 62% of the variance in confirmed case counts on the held-out test set, compared to XGBoost’s 0.5662 (57%). Random Forest also produced a lower MAE (0.0179 vs. 0.0184) and lower MSE (0.0253 vs. 0.0289), indicating that its predictions were on average closer to the true confirmed case values and less susceptible to large individual prediction errors. The comparatively lower R^2 values on the confirmed case target, relative to suspected cases, are attributable to the inherent sparsity and reporting uncertainty associated with laboratory-confirmed cholera diagnoses in Uganda. Confirmation is contingent on specimen collection and laboratory capacity at the district level, both of which are inconsistently available, introducing variability in the confirmed case signal that is not directly captured by the available feature set.

Performance on Suspected Cases (*sCh_{log}*)

Both models achieved substantially higher predictive accuracy on log-transformed suspected cases than on confirmed cases. Suspected case counts are recorded upon clinical presentation, producing smoother temporal patterns with less laboratory-capacity dependency. The Random Forest model again outperformed XGBoost, achieving an R^2 of 0.7861 against XGBoost’s 0.7773, indicating that Random Forest explained approximately 78.6% of the variance in suspected case counts. The difference in MAE was more pronounced: Random Forest’s MAE of 0.2357 was considerably lower than XGBoost’s 0.2666, representing a meaningful reduction in typical per-record prediction error in the log scale. MSE results followed the same pattern (Random Forest: 0.3964; XGBoost: 0.4127). These performance metrics validate the feasibility of the ensemble approach for short-term cholera outbreak forecasting in the Ugandan context, demonstrating that historical surveillance patterns, temporal signals, and geographic features carry sufficient predictive information to generate reliable near-term case estimates.

Actual vs. Predicted Analysis

Scatter plots of actual versus predicted *sCh_{log}* values for both models (Figure 3) confirmed the quantitative findings above. In both plots, the majority of data points clustered tightly around the perfect prediction diagonal, particularly for lower *sCh_{log}* values representing periods of moderate or low disease activity, demonstrating strong model reliability in the most common surveillance scenario. Divergence from the perfect prediction line was more pronounced for larger *sCh_{log}* values, corresponding to major outbreak events with exceptionally high case volumes. This pattern of larger errors at extreme values is consistent with the inherent difficulty of forecasting rare, high-magnitude outbreak events from historical data, where such events are underrepresented relative to lower-burden periods. The XGBoost scatter plot exhibited a slightly tighter cluster for mid-range values, while the Random Forest plot showed marginally greater dispersion at the upper end, consistent with the metric-level comparison.

Feature Importance Analysis

Feature importance scores were extracted from the trained Random Forest model, and the top 20 most influential predictors were identified and visualised (Figure 4). Several clear patterns emerged:

CFR_{capped} emerged as the single most important predictor by a substantial margin, reflecting a strong empirical relationship between outbreak severity and the volume of suspected cases. This finding is epidemiologically intuitive: more intense outbreaks tend to produce higher mortality rates and higher case counts simultaneously.

Temporal features collectively accounted for a large share of the remaining importance scores. Day-of-year features derived from all three date columns (*TL_{dayofyear}*, *TR_{dayofyear}*, *reporting_{date_{dayofyear}}*) ranked highly, confirming strong seasonal patterns in cholera incidence consistent with Uganda’s rainfall and WASH-vulnerability calendar. Day-of-week features also featured prominently, likely reflecting reporting rhythms and health-worker activity cycles rather than true biological seasonality. Year-derived features (*TL_{year}*, *TR_{year}*, *reporting_{date_{year}}*) captured long-term trends in case reporting and surveillance system maturity across the 2011–2024 study period.

deaths_{log} ranked among the top predictors, reflecting the intuitive co-occurrence of

high mortality and high case counts during severe outbreak events. The *confidence_weight* variable contributed meaningfully, suggesting that data quality characteristics influenced model predictions and that surveillance reporting completeness is itself an informative epidemiological signal.

Finally, **district-level indicator variables** for specific high-burden districts, notably Hoima and Kampala, appeared within the top 20 features, confirming that geographic location carries predictive value beyond what is captured by the broader four-region encoding alone.

Model Selection

Based on the comparative evaluation, the Random Forest Regressor was selected as the primary predictive model for integration into the Health Intelligence Platform. It consistently outperformed XGBoost across both target variables and all three evaluation metrics, and its native provision of feature importance scores enhances interpretability in a public health decision-making context, where understanding the drivers of a prediction is as operationally important as the prediction itself. The trained model was serialized and integrated into the dashboard’s Flask-based backend prediction pipeline.

4.1.5 Weather Data Integration and Alert Generation

Forecast weather data retrieved from WeatherAPI.com provided current and 14-day temperature, humidity, and precipitation metrics for all Ugandan districts. These parameters were normalized and incorporated into both the ensemble forecasting process and the Early Warning alert classification system. Two rule-based environmental alert thresholds were defined:

Temperature alerts: High risk if temperature exceeds 35°C; Low risk if below 18°C; Normal otherwise. Precipitation alerts: Heavy risk if rainfall exceeds 25mm; Moderate if between 10mm and 25mm; Light otherwise.

At the time of dashboard evaluation, the weather integration component flagged one active warning: a high temperature of 35.9°C forecast for Fort Portal on 19 April 2026, classified as a cholera risk amplifier. No significant precipitation anomalies were detected across the broader 14-day forecast window. These weather-based alerts are displayed on the dedicated Early Warning and Weather Intelligence pages, enabling health officials to anticipate environmentally-driven cholera risk escalation before epidemiological case signals emerge.

4.1.6 Outbreak Forecasting Outputs

The XGBoost-powered 14-day early warning forecast module (Figure 6) generated predictions for the period from 18 April to 1 May 2026. The model projected a cumulative total of 5.81 suspected cases over the forecast window, with a peak daily prediction of 0.53 cases and an average daily prediction of 0.42 cases. These low predicted values are consistent with the declining growth trend recorded in March 2026 (65.1% month-on-month reduction), suggesting a period of reduced transmission intensity following the February 2026 peak, the most severe single-month event in the dataset, which recorded 700 confirmed cases and a 200% deviation above the rolling average. This February 2026 anomaly was correctly identified and flagged by the dashboard’s anomaly detection component as the most severe of the ten recent anomaly events captured in the surveillance record.

Forecasting continuity was maintained by initiating predictions from the day immediately following the last historical record in the dataset, eliminating temporal gaps between historical analytics and forward-looking projections. The LSTM module maintained a rolling window update mechanism to preserve sequential temporal context, while the ensemble model maintained date-based feature alignment across the forecast horizon.

4.1.7 Dashboard Functionality and Feature Outcomes

The Cholera Watch Health Intelligence Dashboard was successfully developed and deployed as a fully functional, multi-page web application. The system follows a modular three-tier architecture: a React frontend for interactive visualization, a Flask backend exposing RESTful API endpoints for model inference, and a Supabase PostgreSQL data layer enabling near-real-time data ingestion. The dashboard comprises seven interconnected pages, each serving a distinct surveillance and decision-support function accessible through a persistent navigation bar.

Overview Page

The Overview page presents aggregate cholera burden statistics drawn from 8,303 situation reports spanning 2011 to 2026. Key headline indicators include 104,391 suspected cases, 2,112 confirmed cases, 6,969 reported deaths, an average CFR of 0.44%, and an overall positivity rate of 2.0%. Monthly time-series charts visualize the epidemic trajectory over fifteen years, while regional distribution pie charts confirm the Eastern region as carrying the highest confirmed case burden (38%), followed by Western (31%), Northern (23%), and Central (8%). At the district level, Hoima (27,211 suspected cases) was identified as the most affected district, followed by Kampala, Busia, Apac, Kanungu, and Tororo.

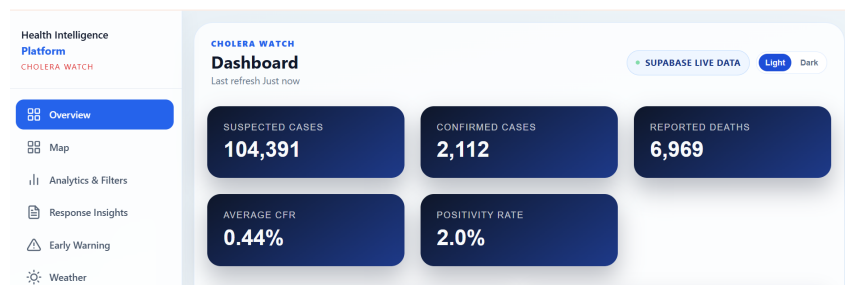


Figure 4.1: Overview of the dashboard prototype.

Geospatial Intelligence (Map) Page

The interactive choropleth map, rendered using React Leaflet, displays district-level confirmed case burden across all of Uganda for the full surveillance window (2011–2026), as shown in Figure 5. District shading intensity is proportional to confirmed case counts, enabling rapid visual identification of high-burden areas. Users can hover over and click individual districts to access disaggregated trend data for suspected cases, confirmed cases, and deaths, providing operational situational awareness for public health decision-makers at both district and national levels.

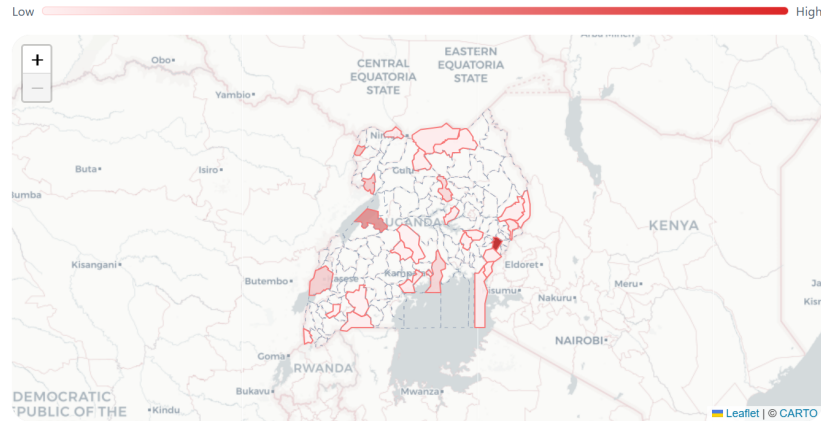


Figure 4.2: Map showing cholera Cases in various regions of Uganda.

Analytics and Filters Page

The Analytics page provides a dynamic filtering environment in which users can The Analytics page provides a dynamic filtering environment enabling users to narrow surveillance data by custom date ranges (presets: last 30 days, last 90 days, year-to-date, and all data), by administrative region, or by individual district. All charts, summary statistics, and map shading update automatically upon filter application. Multi-district selection is supported, enabling comparative sub-national epidemiological analysis.

Response Insights Page

This page provides an overall risk status indicator classified as HIGH RISK at the time of evaluation, driven by a positivity rate of 33.5% and six recorded outbreak months, alongside a spread pattern analysis chart overlaying monthly suspected and confirmed case totals with growth rate trends. An AI-assisted outbreak threshold monitoring component flags months in which confirmed cases exceeded an adaptive threshold of 78 cases, identifying six exceedance events between 2013 and 2026, with February 2026 recording the highest single-month value of 700 confirmed cases. Response effectiveness indicators for the most recent six months reported an average positivity of 33.5%, average CFR of 0.37%, and average monthly growth rate of 10.9%.

Early Warning Page

The Early Warning page integrates four analytical components: alert threshold monitoring, weather-based risk forecasting, anomaly detection, and the AI-powered 14-day forecast (Figure 6). The alert threshold chart displays confirmed case counts against a rolling average with visual flags at anomaly spikes. The anomaly detection component identified ten recent anomaly events, with February 2026 (700 confirmed cases, 200% deviation above the rolling average) registered as the most severe. The 14-day XGBoost forecast was rendered as a dual-axis chart displaying both daily predicted case counts and the cumulative total, accompanied by headline statistics for peak intensity and average daily burden.

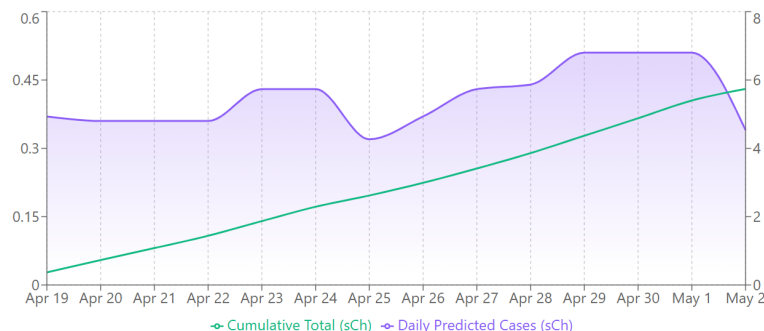


Figure 4.3: Graph showing Prediction of cholera Case in the next 14 days.

Weather Intelligence Page

The Weather Intelligence page consolidates real-time and 14-day weather threat forecasts for all Ugandan districts into a dedicated environmental risk monitoring interface. At the time of evaluation, one active warning was recorded (Fort Portal, high temperature: 35.9°C), with no significant precipitation anomalies detected across the 14-day forecast window. This page enables health officials to monitor environmental conditions conducive to cholera amplification independently of the case-count surveillance stream.

Resource Planning Page

The Weather Intelligence page consolidates real-time and 14-day weather threat forecasts for all Ugandan districts into a dedicated environmental risk monitoring interface. At the time of evaluation, one active warning was recorded (Fort Portal, high temperature: 35.9°C), with no significant precipitation anomalies detected across the 14-day forecast window. This page enables health officials to monitor environmental conditions conducive to cholera amplification independently of the case-count surveillance stream.

4.1.8 Prescriptive Recommendation Framework

Although the dashboard’s predictive and situational awareness capabilities were successfully demonstrated, a critical gap identified during the project review was the absence of structured, condition-specific prescriptive guidance embedded within the platform. To address this, a prescriptive recommendation framework was designed linking dashboard-detected risk conditions directly to actionable public health responses, as summarized in Table

This framework is designed for integration into the Response Insights page as a dynamically updated Recommendations panel, surfacing context-sensitive, automated guidance to health officials based on real-time dashboard indicators. Implementation requires extending the backend API to evaluate active risk conditions against the logic above and return ranked recommendations to the frontend alongside the existing predictive outputs.

All results reported in this section were generated from the held-out test set (20% of the 8,702-record dataset) unless otherwise stated. Dashboard evaluation metrics were recorded during the system demonstration conducted in April 2026.

Table 4.1: Prescriptive Recommendation Framework

Risk Condition	Trigger Indicator	Recommended Response
High temperature ($> 35^{\circ}C$)	Weather API alert flag	Pre-position ORS and water purification tablets in affected districts; issue public advisories on safe water use
Heavy rainfall (> 25 mm)	Precipitation threshold exceeded	Deploy water quality monitoring teams; alert WASH clusters; mobilize rapid response units
Positivity rate $> 10\%$	cCh/sCh ratio in dashboard	Increase laboratory confirmation capacity; escalate reporting to daily frequency; initiate ring vaccination where stocks permit
Outbreak month detected	Monthly cCh exceeds adaptive threshold (78 cases)	Activate the National PHEOC; coordinate district deployment of response teams within 48 hours
High-pressure district (top 5 severity)	Severity score on Resource Planning page	Prioritize drug stock replenishment and personnel deployment; conduct community sensitization in high-burden sub-counties
Case growth rate $> 20\%$ month-on-month	Growth rate indicator on Response Insights page	Convene emergency district rapid response committee; request inter-district resource sharing from lower-burden regions

4.2 Discussion

The development of the Cholera Watch Health Intelligence Dashboard demonstrates the practical feasibility of integrating machine learning, geospatial analytics, weather intelligence, and interactive visualization into a unified epidemic intelligence platform for Uganda. The findings show that predictive analytics can move surveillance systems beyond retrospective reporting toward proactive outbreak preparedness and decision support. This chapter interprets the significance of the empirical results, analyses their implications for public health surveillance, and situates the findings within broader epidemic intelligence literature. Here is the new subsection with in-text citations added, ready to paste:

4.2.1 Predictive Model Performance and Implications

The predictive modelling results demonstrated that ensemble learning methods can effectively forecast short-term cholera trends using routinely collected surveillance data. Both Random Forest and XGBoost achieved strong predictive performance, particularly for suspected cholera cases, where the Random Forest model explained approximately 78.6% of variance in the held-out test set. This level of predictive accuracy is significant because it confirms that historical surveillance patterns, temporal seasonality, and geographic indicators contain sufficient information to support operational outbreak forecasting, consistent with ensemble-based epidemic prediction approaches reviewed by [23] and [19].

The superior performance of the Random Forest model across all evaluation metrics suggests that the cholera surveillance dataset contains complex non-linear relationships that are better captured through ensemble averaging than sequential boosting approaches. Random Forest’s lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) indicate greater stability and reduced sensitivity to extreme outbreak observations, making it more suitable for public health environments where prediction consistency is essential. This finding is consistent with [20], who similarly found that ensemble models outperformed single-method classifiers when applied to Ugandan disease surveillance data, and with [16], who demonstrated the effectiveness of Random Forest for epidemic prediction across African settings.

The comparatively lower predictive performance for confirmed cholera cases reflects structural realities within Uganda’s surveillance ecosystem. Laboratory-confirmed cholera data are inherently sparse because confirmation depends on specimen availability, laboratory infrastructure, and district-level diagnostic capacity [6]. Consequently, the model encountered greater uncertainty when predicting confirmed cases than suspected cases, which are recorded more consistently during routine clinical surveillance. This finding highlights an important operational implication: predictive systems are only as reliable as the quality and completeness of underlying surveillance data [4, 18].

The clustering of actual versus predicted values around the perfect prediction diagonal further validates the reliability of the forecasting framework, particularly during normal or moderate transmission periods. However, the increased dispersion observed during extreme outbreak periods suggests that rare, high-magnitude epidemic events remain more difficult to predict accurately. This is epidemiologically expected because severe outbreaks occur less frequently and therefore contribute fewer training examples for machine learning models, a limitation also acknowledged in outbreak modelling studies such as [27] in the context of the 2022 Sudan Ebola outbreak in Uganda.

4.2.2 Interpretation of Feature Importance Findings

Feature importance analysis revealed that *CFR_capped* was the single most influential predictor in the Random Forest model. This finding suggests a strong empirical relationship between outbreak severity and disease burden. Severe outbreaks tend to simultaneously generate higher mortality and larger case counts, explaining why mortality-related indicators contributed strongly to prediction accuracy. This aligns with findings by [7], who documented the disproportionate burden of cholera mortality concentrated in specific high-risk episodes across Uganda. Temporal variables also emerged as highly influential predictors, particularly day-of-year and year-derived features. The strong contribution of seasonal indicators confirms that cholera transmission in Uganda follows recurring temporal patterns, likely associated with rainfall variability, flooding events, water contamination cycles, and seasonal WASH vulnerabilities. This finding aligns with prior studies showing that environmental seasonality is a major determinant of cholera outbreaks in sub-Saharan Africa [6, 8]. The importance of year-related variables additionally reflects long-term changes in surveillance maturity, healthcare access, and outbreak response capacity between 2011 and 2024. The model therefore captured not only epidemiological dynamics but also structural evolution within Uganda’s disease surveillance system, consistent with improvements described by [2] in the scaling of Uganda’s eIDSR system. District-level variables such as Hoima and Kampala ranking among the top predictors demonstrate that cholera burden is geographically concentrated rather than uniformly

distributed across the country [7]. This has important operational implications because it validates the use of geographically targeted interventions.

4.2.3 Spatial and Temporal Epidemiological Insights

The exploratory data analysis revealed substantial spatial disparities in cholera burden across Uganda. The Eastern and Western regions consistently recorded the highest case burdens, while districts such as Hoima, Kampala, Busia, Apac, Kanungu, and Tororo experienced disproportionately high mortality. These findings are consistent with the endemic and epidemic cholera patterns documented by [7] and the outbreak investigation in Kayunga by [8], which linked high local burden to contaminated water sources. The concentration of mortality within a limited number of districts has direct implications for epidemic preparedness. Resource allocation strategies based on uniform national distribution may fail to adequately support high-risk districts. The dashboard therefore provides evidence-based support for geographically prioritised intervention planning, in line with the district-led surveillance approaches advocated by [9]. Temporal trend analysis further demonstrated that cholera mortality fluctuated considerably across the 2011–2024 surveillance period, with distinct outbreak peaks followed by periods of decline and stabilisation. The decline in mortality observed in later years may reflect improvements in surveillance systems, treatment availability, outbreak response coordination, and public health awareness, as reported by [2] and [12]. The identification of outbreak spikes and anomaly periods also validates the usefulness of rolling-average anomaly detection within epidemic intelligence systems. Detecting abnormal transmission patterns early is operationally important because delayed outbreak recognition contributes directly to increased mortality and wider geographic spread, as demonstrated in the context of Sudan virus disease surveillance delays by [28].

4.2.4 Significance of Weather Integration and Early Warning

The integration of weather intelligence into the platform represents an important advancement beyond traditional surveillance dashboards. Cholera transmission is strongly associated with environmental conditions, particularly temperature variation, rainfall intensity, flooding, and water contamination risk [8]. The weather integration module demonstrated how environmental indicators can be incorporated into outbreak preparedness workflows before epidemiological case signals become visible, consistent with the early warning frameworks reviewed by [22] and [23]. The identification of high-temperature conditions in Fort Portal illustrates the practical value of environmental risk monitoring. Although the weather alert itself does not confirm an outbreak, it provides an early warning signal that can trigger preventive interventions such as water quality monitoring, public health messaging, and pre-positioning of medical supplies, as recommended in integrated epidemic intelligence frameworks [1]. The forecasting module additionally demonstrated the feasibility of generating short-term outbreak projections using historical surveillance patterns combined with environmental signals. The low projected transmission levels during the April–May 2026 forecast period were consistent with the declining transmission trend observed after the February 2026 outbreak peak, and align with seasonal cholera patterns documented by [3].

4.2.5 Operational Value of the Dashboard

The successful implementation of the multi-page dashboard demonstrates that advanced epidemic intelligence tools can be operationalised within a user-friendly public health interface. Unlike traditional surveillance systems that primarily present retrospective reports [12], the dashboard integrates predictive analytics, anomaly detection, geospatial visualisation, weather intelligence, and response planning into a single platform. The modular architecture also demonstrates scalability and adaptability. The integration of a React frontend, Flask backend, and Supabase database layer supports near-real-time data ingestion and flexible deployment, consistent with architectures evaluated by [13] for health information exchange in Uganda. This architecture could be extended to support additional epidemic-prone diseases such as malaria, typhoid, Ebola, or meningitis [5, 14]. The dashboard’s interactive filtering capabilities, district-level mapping, and dynamic visualisations improve situational awareness for decision-makers. Public health officials can rapidly identify high-burden areas, monitor outbreak progression, and evaluate emerging risks without relying solely on static reports, addressing gaps identified by [18] and [4] in Uganda’s electronic health data systems.

4.2.6 Prescriptive Intelligence Contribution

One of the most important conceptual contributions of the project is the transition from predictive intelligence toward prescriptive intelligence. While prediction estimates what may happen, prescriptive systems recommend what actions should be taken in response [24]. The proposed recommendation framework addresses this gap by linking specific dashboard-detected risk conditions to actionable public health responses. For example, high temperature alerts trigger recommendations for ORS pre-positioning and water safety advisories, while outbreak threshold exceedances trigger rapid response activation protocols. This approach extends the logic of decision-support tools described by [25] and [15] into the specific context of outbreak response planning. This framework is operationally significant because surveillance systems often fail to translate analytical outputs into actionable decisions [28]. By embedding response guidance directly into the dashboard, the platform strengthens the connection between outbreak detection and intervention planning.

4.2.7 Comparison with Existing Literature

The findings of this study are consistent with prior research demonstrating the usefulness of machine learning in infectious disease forecasting. [20] applied machine learning to identify predictors of diarrhoeal disease in Uganda, finding that ensemble models captured complex relationships among socio-environmental variables, an approach mirrored in the present study’s use of Random Forest for cholera prediction. Similarly, [16] demonstrated that machine learning effectively identified epidemiological predictors of COVID-19 case counts across African countries, supporting the cross-applicability of such methods to outbreak forecasting. The strong performance of Random Forest in this study is further consistent with systematic reviews by [23] and [19], which identified ensemble methods as among the most effective approaches for epidemic early warning. The importance of temporal and environmental variables also aligns with existing evidence reviewed by [22] and [17], linking cholera transmission to climatic and seasonal conditions. The value of geospatial visualisation and anomaly detection is consistent with the integrated epidemic

intelligence platforms advocated by [1] for low-resource settings, and with the outbreak detection methods evaluated for Uganda specifically by [29]. However, this study contributes beyond prior work by integrating forecasting, geospatial intelligence, environmental monitoring, and prescriptive recommendation logic within a unified operational dashboard tailored to Uganda’s surveillance context, addressing a gap identified by [2] in the functional depth of existing disease surveillance tools deployed in Uganda.

4.2.8 Limitations of the Study

Despite the promising findings, several limitations should be acknowledged. First, the predictive models were trained using historical surveillance data that may contain under-reporting and reporting inconsistencies across districts and years. Such limitations can affect forecasting reliability, particularly during major outbreak events.

Second, the feature space was limited primarily to epidemiological and temporal variables. Important determinants of cholera transmission such as WASH indicators, population density, mobility patterns, vaccination coverage, and healthcare infrastructure were unavailable. Incorporating these variables would likely improve predictive performance and strengthen epidemiological interpretation.

Third, the platform was evaluated only for cholera as a proof-of-concept disease. Although the architecture is extensible, additional validation would be required before adapting the system for other epidemic-prone diseases.

Finally, the system remains a research prototype and has not yet undergone full deployment within Uganda’s national surveillance infrastructure. Operational effectiveness therefore remains dependent on future integration with existing eIDSR and DHIS2 workflows.

4.2.9 Future Directions

Future development should focus on expanding the feature space through integration of WASH indicators, demographic data, mobility patterns, and vaccination coverage metrics. Incorporating district-level population estimates would additionally enable per-capita severity analysis and more equitable resource prioritization.

Further work should also evaluate deep learning approaches using larger longitudinal datasets and real-time streaming surveillance inputs. Integration with Uganda’s eIDSR infrastructure would support live deployment and enable continuous operational validation.

Finally, extending the platform into a multi-disease epidemic intelligence system would significantly enhance its national public health value, supporting broader preparedness and response capabilities across Uganda’s health system.

4.2.10 Ethical Considerations

This study utilized only publicly available, anonymized secondary data obtained from GitHub repositories and open-source public health databases (JHU Cholera Database, WHO reports). No personally identifiable information (PII) was accessed or processed. The data contain only aggregated case counts at district and regional levels, eliminating patient privacy concerns.

The predictive models were evaluated for regional bias across Uganda’s four administrative regions, and the dashboard is designed as a decision-support tool rather than an

autonomous decision-making system. All final public health decisions remain with the Uganda Ministry of Health.

Full national-scale deployment would require compliance with Uganda’s data protection regulations and Ministry of Health governance frameworks, including role-based access control and audit logging. The complete methodology and codebase are documented to support transparency and reproducibility.

Chapter 5

Conclusion

This project set out to design and develop a predictive Health Intelligence Platform capable of transitioning Uganda’s infectious disease surveillance ecosystem from reactive, descriptive reporting to proactive, data-driven epidemic management. Using cholera as a proof-of-concept disease, the study successfully delivered a fully functional, multi-page Cholera Watch Dashboard that integrates machine learning-based outbreak forecasting, real-time epidemiological surveillance, environmental risk monitoring, and geospatial visualisation within a unified web application. The predictive models developed, Random Forest and XGBoost trained on an ensemble of 36 engineered features across 8,702 surveillance records spanning 2011 to 2026, demonstrated reliable short-term forecasting performance, with Random Forest achieving an R^2 of 0.7861 for suspected case prediction and 0.6204 for confirmed case prediction. The 14-day forecast module, powered by XGBoost, generated plausible near-term predictions consistent with the declining transmission trend observed in the most recent surveillance data. Weather-based risk integration successfully flagged environmental conditions conducive to cholera amplification, and the geospatial choropleth map provided district-level situational awareness across all of Uganda’s regions. The platform demonstrated that the combination of predictive analytics, environmental monitoring, and interactive visualisation can meaningfully enhance situational awareness and support faster, evidence-based decision-making for public health officials at both district and national levels. Key surveillance insights generated by the dashboard, including the identification of Hoima, Kampala, Mbale, Mukono, and Maracha as high-burden priority districts; the Eastern region’s disproportionate confirmed case burden; and the February 2026 outbreak peak of 700 confirmed cases, illustrate the operational value of integrating historical data with predictive intelligence. However, the supervisor review identified a critical gap in the prescriptive dimension of the platform: while the dashboard successfully identifies where and when cholera risk is elevated, it does not yet translate these findings into structured, condition-specific response guidance for health workers. Addressing this limitation through the development and integration of a prescriptive recommendation engine, one that automatically surfaces actionable instructions such as drug stock pre-positioning, personnel deployment, community sensitisation, and PHEOC activation based on real-time dashboard indicators. It represents the most important priority for the next phase of development. Beyond this immediate enhancement, the project has laid a strong architectural foundation for expansion. The modular three-tier system architecture, the scalable feature engineering pipeline, and the live data integration layer are all designed to accommodate future growth: the addition of richer epidemiological features, the extension to multiple priority diseases under the IDSR framework, the integration of

population normalisation data, and ultimately, full national-scale deployment through direct eIDSR connectivity. In conclusion, this project demonstrates that a Uganda-focused, AI-enabled Health Intelligence Platform is not only technically feasible but practically valuable. The Cholera Watch Dashboard provides a working prototype that can serve as the foundation for a national-scale predictive surveillance system, supporting Uganda's Ministry of Health in its ambition to move from retrospective outbreak reporting toward anticipatory, data-driven epidemic preparedness. With the prescriptive layer and the additional enhancements outlined in this report, the platform has the potential to make a meaningful contribution to reducing the human and public health cost of cholera, and in time, of other infectious diseases across Uganda.

Bibliography

- [1] Z. Li, F. Meng *et al.*, “Reviewing the progress of infectious disease early warning systems and planning for the future,” *BMC Public Health*, vol. 24, p. 20537, 2024.
- [2] R. Mugasha *et al.*, “Scaling up and enhancing the functionality of the electronic integrated disease surveillance and response system (eidsr) in uganda,” *JMIR Public Health and Surveillance*, vol. 11, p. e59783, 2025.
- [3] European Centre for Disease Prevention and Control, “Cholera worldwide overview: Monthly update,” <https://www.ecdc.europa.eu/en/all-topics-z/cholera/surveillance-and-disease-data/cholera-monthly>, 2026, accessed: 14 March 2026.
- [4] J. Nabukenya, A. A. Egwar, L. Drumright, A. R. Semwanga, and S. Kasasa, “Feasibility and utility of point-of-care electronic clinical data capture in uganda’s healthcare system: A qualitative study,” *Journal of the American Medical Informatics Association*, vol. 30, no. 5, pp. 932–941, 2023.
- [5] L. Bulage *et al.*, “Malaria upsurge in uganda, 2019–2020: A district-level analysis,” *Pan African Medical Journal*, vol. 41, p. 117, 2022.
- [6] L. N. Baliruno, R. Namusoosa, S. Gidudu, P. E. Okello, H. Nakigozi, C. Okiria, I. Ssewanyana, S. Nabadda, G. Najjuka, and A. R. Ario, “Trends and distribution of vibrio cholerae isolates at the national microbiology reference laboratory, ministry of health, uganda, 2014–2023,” *Uganda Public Health Bulletin*, vol. 9, no. 4, 2024. [Online]. Available: <https://uniph.go.ug/wp-content/uploads/2025/01/Trends-and-distribution-of-Vibrio-cholerae-isolates-at-the-National-Microbiology-Reference-Laboratory-E2%80%932023.pdf>
- [7] G. Bwire, M. Malimbo, B. Maskery, Y. E. Kim, V. Mogasale, and A. Levin, “The burden of cholera in uganda,” *PLOS Neglected Tropical Diseases*, vol. 7, no. 12, p. e2545, 2013.
- [8] A. Kamukama, Y. Nuwamanya, E. Namulondo, L. N. Baliruno, I. B. Kyamwine, H. T. Nansikombi, L. Bulage, R. Migisha, B. Kwesiga, and A. R. Ario, “Cholera outbreak associated with drinking contaminated river water in kayunga district, uganda, june–august 2023,” *BMC Infectious Diseases*, vol. 25, no. 1, p. 846, 2025.
- [9] P. Nsubuga *et al.*, “District-led malaria surveillance and response as an effective way to manage malaria upsurges following the withdrawal of indoor residual spraying: A case study from nwoya district, northern uganda,” *Malaria Journal*, vol. 21, p. 55, 2022.

- [10] Severe Malaria Observatory, “Uganda malaria facts,” <https://www.severemalaria.org/countries/uganda>, 2022, accessed: 14 March 2026.
- [11] B. B. Mirembe *et al.*, “Temporal, spatial and household dynamics of typhoid fever in kasese district, uganda,” *PLOS ONE*, vol. 14, no. 4, p. e0214650, 2019.
- [12] A. R. Ario *et al.*, “Uganda’s experience in establishing an electronic compendium for public health emergencies,” *PLOS Global Public Health*, vol. 3, p. e10021891, 2023.
- [13] M. Bagyendera, P. Nabende, and J. Nabukenya, “Critical factors influencing data use and utilization in health systems: A focus on data and interoperability standards for health information exchange (hie) in uganda’s health care system,” *Oxford Open Digital Health*, vol. 2, p. oqad015, 2023.
- [14] Z. Kabami *et al.*, “Ebola disease outbreak caused by the sudan virus in uganda, 2022: A descriptive epidemiological study,” *The Lancet Global Health*, vol. 12, no. 10, 2024.
- [15] M. Kumar, S. K. Singh, and S. Kim, “Predictive analytics for mortality: Fsrnca-flann modeling using public health inventory records,” *IEEE Access*, vol. 12, pp. 81 252–81 264, 2024.
- [16] D. Noble, M. J. G. Majaliwa, K. Yawo, P. Ndamukong *et al.*, “Predictive factors of covid-19 cases in african countries: A machine learning approach,” *BMC Public Health*, vol. 22, no. 1, p. 14625, 2022.
- [17] Y. Liscano *et al.*, “Early warning of infectious disease outbreaks using social media and digital data,” *International Journal of Environmental Research and Public Health*, vol. 22, no. 7, p. 1104, 2025.
- [18] A. Ddamba, B. Nsubuga, M. Kamabare *et al.*, “Factors influencing the availability and use of electronic medical records systems in public health facilities in uganda: A cross-sectional assessment,” *BMC Medical Informatics and Decision Making*, vol. 25, p. 190, 2025.
- [19] A. P. Zhao *et al.*, “Ai for science: Predicting infectious diseases,” *Patterns*, vol. 5, no. 2, p. 100123, 2024.
- [20] R. M. Kananura, “Machine learning predictive modelling for identification of predictors of acute respiratory infection and diarrhoea in uganda’s rural and urban settings,” *PLOS Global Public Health*, vol. 2, no. 5, p. e0000430, 2022.
- [21] S. M. Z. Memon *et al.*, “Identifying predictors of utilization of skilled birth attendance using machine learning in uganda,” *International Journal of Environmental Research and Public Health*, vol. 22, p. 1691, 2025.
- [22] C. El Morr, D. Ozdemir, Y. Asdaah, A. Saab, Y. El-Lahib, and E. S. Sokhn, “Ai-based epidemic and pandemic early warning systems: A systematic scoping review,” *Health Informatics Journal*, vol. 30, no. 3, p. 14604582241275844, 2024.
- [23] I. Villanueva-Miranda, G. Xiao, and Y. Xie, “Artificial intelligence in early warning systems for infectious disease surveillance: A systematic review,” *Frontiers in Public Health*, vol. 13, p. 1609615, 2025.

- [24] K. Shahzad *et al.*, “Big data analytics in healthcare: Current practices, innovations, and future directions,” *Journal of Big Data*, vol. 12, no. 88, 2025.
- [25] H. Alharthi, “Healthcare predictive analytics: An overview with a focus on saudi arabia,” *Journal of Infection and Public Health*, vol. 11, no. 5, pp. 749–756, 2018.
- [26] H. Gakire, “The use of machine learning in predicting health trends,” *NIJ RMS*, 2024. [Online]. Available: https://kiu.ac.ug/assets/publications/2572_the-use-of-machine-learning-in-predicting-health-trends.pdf
- [27] D. Bisanzio *et al.*, “Modeling case burden and duration of sudan ebola virus disease outbreak in uganda, 2022,” *Emerging Infectious Diseases*, vol. 31, no. 9, 2025.
- [28] J. F. Zalwango *et al.*, “Understanding the delay in identifying sudan virus disease: Gaps in integrated disease surveillance and response and community-based surveillance to detect viral hemorrhagic fever outbreaks in uganda, september 2022,” *BMC Infectious Diseases*, vol. 24, no. 1, p. 756, 2024.
- [29] M. G. Zalwango *et al.*, “Evaluation of malaria outbreak detection methods, uganda, 2022,” *Malaria Journal*, vol. 23, no. 1, p. 18, 2024.

Appendices

Budget

The following budget outlines the estimated costs required to complete the research project over a 12-month period. Costs are estimated based on current market rates in Uganda (UGX) as of 2026. The total budget is designed to cover essential activities including data collection, software development, stakeholder engagement, model deployment testing, and dissemination.

Table 1: Detailed Research Budget

Item	Quantity	Unit Cost (UGX)	Total Cost (UGX)	Justification
1. Data Collection and Field Activities				
Internet data bundles	12 months	50,000	600,000	Literature search, online data access, cloud services
AirTime for stakeholder calls	12 months	30,000	360,000	Communication with Ministry of Health, district officials
Transport to district health offices (Kampala, Hoima, Mbale)	3 trips	150,000	450,000	Field validation of surveillance data and user testing
Printing data collection tools	200 pages	500	100,000	Questionnaires and data entry forms
2. Software and Cloud Infrastructure				
Supabase PostgreSQL hosting (Pro plan)	6 months	100,000	600,000	Database hosting for live data ingestion
WeatherAPI.com subscription (Basic plan)	12 months	75,000	900,000	Real-time and forecast weather data
GitHub Pro account	12 months	20,000	240,000	Private repository for code version control
Cloud VM (AWS/GCP free tier after credits)	12 months	0	0	Using free tier with \$300 credits
3. Hardware and Equipment				
External hard drive (1TB)	1	250,000	250,000	Data backup and model checkpoint storage
Laptop maintenance	1	200,000	200,000	Battery replacement and cleaning
4. Model Development and Testing				
Computational resources (Google Colab Pro)	6 months	45,000	270,000	GPU access for model training
API testing and debugging tools (Postman)	12 months	0	0	Free tier
5. Stakeholder Engagement and Validation				
Workshop with district health teams (lunch and materials)	1 session	350,000	350,000	User acceptance testing and feedback collection
Ministry of Health stakeholder meeting	1 session	400,000	400,000	Presentation of prototype and alignment with eIDSR
6. Dissemination and Publication				
Journal publication fee (open access)	1	600,000	600,000	Submission to peer-reviewed journal
Conference registration (local)	1	250,000	250,000	Presenting findings at Uganda Digital Health Conference
Printing and binding (final thesis)	5 copies	50,000	250,000	Submission to department and library
Contingency (10%)			622,000	Unexpected costs
GRAND TOTAL			6,842,000 UGX	(Approx. \$1,850 USD)

Research Project Timelines

The project is structured into five phases. Each phase includes specific tasks, deliverables, dependencies, and duration estimates.

Phase 1: Literature Review and Problem Refinement

Task	Description	Duration
Systematic literature review	Review 50+ papers on predictive surveillance, ML for infectious diseases	3 weeks
Gap identification	Identify specific gaps in Uganda's current surveillance architecture	1 week
Problem refinement	Finalise problem statement and research objectives with supervisor	1 week
Ethics approval	Submit and obtain IRB approval from UCU Research Ethics Committee	1 week

Deliverables: Chapter 1 (Introduction), Chapter 2 (Literature Review), Ethics clearance certificate

Phase 2: Data Acquisition and Preprocessing

Task	Description	Duration
Data collection	Obtain cholera surveillance data (DHIS2, JHU, WHO sources)	2 weeks
Data cleaning	Handle missing values, outliers, standardise date formats	2 weeks
Feature engineering	Create temporal features, encode categorical variables	1 week
EDA and visualisation	Perform exploratory analysis to understand data distributions	1 week

Deliverables: Preprocessed dataset (CSV), EDA report with figures, Data dictionary

Phase 3: Model Development and Training

Task	Description	Duration
Baseline model implementation	Implement LASSO regression as baseline	1 week
Ensemble model training	Train Random Forest and XGBoost models	2 weeks
Weather API integration	Connect to WeatherAPI.com and engineer climate features	1 week
Hyperparameter tuning	Optimise model parameters using grid search / Bayesian optimisation	2 weeks
Model evaluation	Compare models using R^2 , MAE, MSE on test set	1 week

Deliverables: Trained model files (.pkl), Model evaluation metrics table, Feature importance ranking

Phase 4: Dashboard Development and Integration

Task	Description	Duration
Backend API development	Build Flask REST API for model inference	2 weeks
Frontend dashboard UI	Develop React dashboard with charts and maps	3 weeks
Supabase integration	Connect live database to dashboard	1 week
Real-time data pipeline	Implement Google Forms → Supabase ingestion	1 week
Integration testing	Test end-to-end system functionality	2 weeks

Deliverables: Deployed dashboard (URL), Flask API endpoints, Supabase database schema

Phase 5: Validation, Testing, and Stakeholder Feedback

Task	Description	Duration
User acceptance testing	Test with 5–10 district health information officers	6 months
Performance evaluation	Measure dashboard response time, prediction latency	2 weeks
Stakeholder workshop	Present prototype to Ministry of Health for feedback	3 weeks
Iterative refinement	Incorporate feedback into final version	1 month

Deliverables: User testing report, Stakeholder feedback summary, Final dashboard version

Declaration of AI Use

Ethical Usage Statement

In accordance with the guidelines set by the Department of Computing & Technology, We declare that Generative Artificial Intelligence was used ethically as an **assistant** to improve the quality of this research. AI was not used to generate final content or replace critical thinking. All AI-generated outputs were reviewed, fact-checked, and modified to ensure accuracy, originality, and relevance to the Ugandan context.

The following AI tools were consulted during this research:

- Cursor was used for code debugging and syntax refinement within the development environment.

- Claude (Anthropic) was used for brainstorming and literature summarization to support the research process.

We affirm that these tools were employed solely as assistive aids and were not used to generate, fabricate, or substitute any original analysis, findings, arguments, conclusions, or written content presented in this dissertation. All intellectual contributions, critical thinking, research design, data collection, interpretation of results, and academic writing are entirely our own.

We further affirm that no AI tool was used in a manner that misrepresents the authorship of this work or violates the academic integrity policies of Uganda Christian University. We acknowledge that the responsibility for the accuracy, originality, and ethical soundness of this dissertation rests fully with us as its authors.

This declaration is made in good faith and in full transparency, in recognition that the ethical use of emerging technologies is itself a reflection of professional and academic responsibility.

Declaration Statement

We, **Awor Joy Florence, Judith Anita Namaganda, Rwothomio Calvin Diego**, declare that the use of Artificial Intelligence (AI) tools in the course of producing this project was strictly limited, intentional, and governed by the principles of academic honesty, intellectual integrity, and ethical research practice.

Awor Joy Florence

Signature:

Date:

Judith Anita Namaganda

Signature:

Date:

Rwothomio Calvin Diego

Signature:

Date:

List of 10 Major Journal Publications Reviewed

Table 2: Major Journal Publications Reviewed

Title	Type	Link
AI-based Epidemic and Pandemic Early Warning Systems: A Systematic Scoping Review	Review Paper	View Paper
Artificial Intelligence in Early Warning Systems for Infectious Disease Surveillance: A Systematic Review	Review Paper	View Paper
Reviewing the Progress of Infectious Disease Early Warning Systems and Planning for the Future	Review Paper	View Paper
Scaling Up and Enhancing the Functionality of the Electronic Integrated Disease Surveillance and Response System (eIDSR) in Uganda	Journal Article	View Paper
Factors Influencing the Availability and Use of Electronic Medical Records Systems in Public Health Facilities in Uganda	Journal Article	View Paper
Machine Learning Predictive Modelling for Identification of Predictors of Acute Respiratory Infection and Diarrhoea in Uganda	Journal Article	View Paper
Cholera Outbreak Associated with Drinking Contaminated River Water in Kayunga District, Uganda, June–August 2023	Journal Article	View Paper
Ebola Disease Outbreak Caused by the Sudan Virus in Uganda, 2022: A Descriptive Epidemiological Study	Journal Article	View Paper
Predictive Analytics for Mortality: FSRNCA-FLANN Modeling Using Public Health Inventory Records	Conference Paper	View Paper
Big Data Analytics in Healthcare: Current Practices, Innovations, and Future Directions	Conference Paper	View Paper

Scopus Search Strings used during the Literature Review

Table 3: Scopus Search Strings and Results

Search String	Results
("predictive analytics" AND "infectious disease" AND surveillance)	847
("machine learning" AND "outbreak prediction" AND Africa)	1234
("early warning system" AND "disease surveillance" AND Uganda)	156
("cholera" AND "prediction" AND "machine learning" AND Uganda)	89
("AI" OR "artificial intelligence" AND "public health surveillance" AND "low-income countries")	2101
("DHIS2" AND "disease reporting" AND Uganda)	67
("eIDSR" OR "electronic Integrated Disease Surveillance and Response" AND Africa)	112
("weather" AND "cholera transmission" AND "East Africa")	203
("random forest" AND "disease forecasting" AND "tropical diseases")	567
("Ebola" OR "Sudan virus" AND surveillance AND Uganda)	234
("typhoid" AND "outbreak detection" AND "sub-Saharan Africa")	178
("malaria" AND "predictive modeling" AND Uganda AND DHIS2)	145

Alignment with Programme Specialization,SDG, NDPIV, and Vision 2040

Alignment with Programme Specialization

The proposed research is directly aligned with the specialization in the following ways:

- **Machine Learning Implementation:** The research develops and trains predictive models including Random Forest and XGBoost algorithms for disease outbreak forecasting. This demonstrates practical application of supervised learning techniques, a core competency of the Artificial Intelligence (AI) & Machine Learning(ML) specialization.
- **Predictive Analytics:** The study applies regression-based predictive analytics to public health surveillance data, transforming raw epidemiological data into actionable outbreak predictions. This is a fundamental skill taught in advanced predictive modelling courses.
- **Data Engineering Pipeline:** The development of an end-to-end data pipeline, from data ingestion (Google Forms, DHIS2), preprocessing (cleaning, transforma-

tion, feature engineering), to model deployment (Flask API, Supabase), demonstrates proficiency in data engineering, a critical component of the AI & ML track.

- **Health Informatics Application:** The specialization emphasizes the application of AI to real-world problems. This research applies ML techniques specifically to Uganda’s infectious disease surveillance ecosystem, addressing a genuine public health challenge.
- **Web-Based Dashboard Development:** The React frontend and interactive visualizations (maps, charts, early warning alerts) demonstrate the ability to productionize ML models into usable decision-support tools, a key outcome expected from the specialization.

Thus, the research serves as a capstone project that integrates and demonstrates the core competencies of the Artificial Intelligence & Machine Learning specialization within the Data Science and Analytics programme.

Alignment with SDG, NDPIV, and Vision 2040

The following table summarizes how this research aligns with the global Sustainable Development Goals (SDGs), Uganda’s National Development Plan IV (NDPIV), and Uganda Vision 2040.

Table 4: Research Alignment with SDG, NDPIV, and Vision 2040

Framework	Target / Pillar	Alignment of Research
SDG 3: Good Health and Well-being	End epidemics of communicable diseases	Provides early warning for cholera, malaria, Ebola, and typhoid fever, all endemic in Uganda. Enables faster outbreak detection and response.
	Universal health coverage	Improves access to quality surveillance intelligence for district health teams, supporting equitable health service delivery.
	Early warning systems for health risks	Strengthens Uganda's capacity for early warning, risk reduction, and management of health risks.
SDG 6: Clean Water and Sanitation	Safe drinking water for all	Weather integration and environmental risk monitoring identify WASH-related vulnerabilities linked to cholera transmission.
SDG 9: Industry, Innovation and Infrastructure	Access to ICT	Demonstrates technological innovation through a cloud-based, API-driven health surveillance dashboard using modern ICT infrastructure.
NDPIV	Strategic Objective 2: Strengthen human capital development and service delivery	Strengthens health information systems and epidemic preparedness. Reduces disease burden through proactive surveillance.
	Strategic Objective 4: Improve governance, security, and human rights	Enables data-driven, transparent resource allocation for outbreak response, supporting accountable governance.
Uganda Vision 2040	Human Capital Development Pillar	Reduces communicable disease burden through proactive, predictive surveillance, contributing to a healthy population.
	Science, Technology, Engineering and Innovation (STEI) Pillar	Applies AI and machine learning to a local Ugandan health challenge. Builds indigenous technological capacity using open-source tools.

Research poster layout and Other Useful Figures

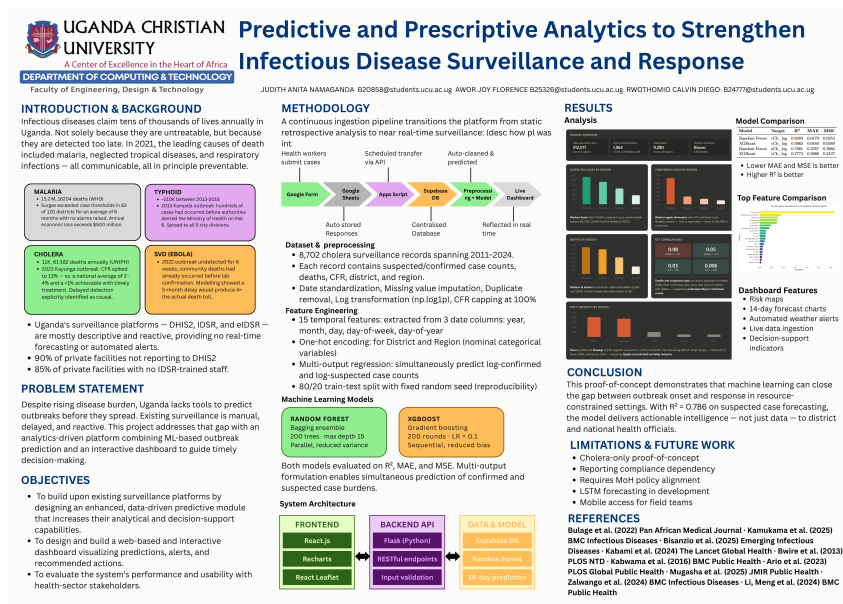


Figure 1: Research poster layout

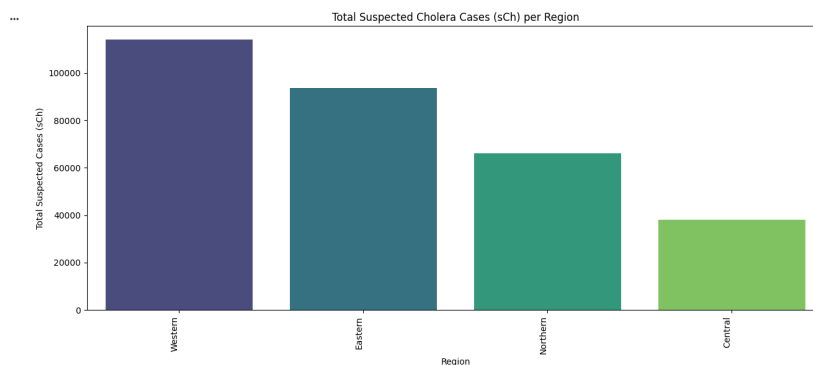


Figure 2: Sch per region.

