MAKERERE UNIVERSITY

A MACHINE LEARNING MODEL FOR PREDICTION OF ANTIBIOTIC RESISTANCE WITH ESCHERICHIA COLI INFECTIONS USING DEMOGRAPHIC, CLINICAL AND MICROBIOLOGICAL DATA

CLARE KAHUMA ALLELUA
2023/HD07/3077U

A DISSERTATION SUBMITTED TO THE SCHOOL OF PUBLIC HEALTH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF HEALTH INFORMATICS AT MAKERERE UNIVERSITY, KAMPALA

January 2026

DECLARATION

I declare that this dissertation is my original work and that it has never been submitted to any institution of higher learning for any award.

…….…… 12/12/2025 Date……...……

Clare Kahuma Allelua

APPROVAL

This is to certify that this dissertation was developed under our supervision and is now ready for submission.

Prof. Noah Kiwanuka

Department of Epidemiology and Biostatistics,

School of Public Health at Makerere University School of Public Health,
Makerere University

……..…….. Date……13/01/2026……

Irene Wanyana,
Assistant Lecturer,
School of Public Health, Makerere University,

……..…….. Date……13/01/2026……

DEDICATION

To my husband (Eng. Paul Nderitu Wanjiku), my father (Eng. Dr. Adolf Kahuma), my mother (Mrs. Kahuma Scholastica) and my siblings, thank you for your love, support, and patience.

ACKNOWLEDGMENTS

I am incredibly grateful for the guidance, encouragement, and support that made this achievement possible. I am especially thankful to my supervisors, Prof. Noah Kiwanuka and Ms. Irene Wanyana, for their patience, expertise, and thoughtful feedback. Their mentorship has been a cornerstone of this work, shaping my skills as a researcher and enriching my understanding of the field.

My sincere thanks also go to the African Center of Excellence in Bioinformatics and Data Intensive Sciences for their essential guidance, especially in the early stages of project development and concept refinement. Additionally, I extend my gratitude to the Global Health Strengthening programme at the Infectious Diseases Institute for providing the datasets that enabled this research.

To my classmates and colleagues, thank you for the companionship and encouragement that carried us through each challenge. Your friendship brought much-needed balance to this academic journey, making it both memorable and rewarding.

I am deeply grateful to my family who instilled in me a strong work ethic and a passion for learning. Your unwavering support and encouragement have been my constant source of strength.

Above all, I thank God for the resilience, wisdom, and guidance throughout this journey. To Him be the glory, now and forever.

TABLE OF CONTENTS.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
AMR	Antimicrobial Resistance
AST	Antimicrobial Susceptibility Testing
E. coli	Escherichia coli
EHRs	Electronic Health Records
ESBLs	Extended-Spectrum Beta-Lactamases
GLASS	Global Antimicrobial Surveillance System
LMICs	Low- and Middle-Income Countries
DR	Drug Resistant
ML	Machine Learning
RRH	Regional Referral Hospital
WISFC	Weighted Importance Score and Frequency Count
WHO	World Health Organization

OPERATIONAL DEFINITION

1. Antimicrobial Resistance (AMR): The ability of microorganisms, such as bacteria, viruses, fungi, and parasites, to resist the effects of medications that once successfully treated infections caused by these organisms. In this study, AMR refers to bacterial resistance to antibiotics.
2. Extended Spectrum Beta-Lactamases (ESBLs): Enzymes produced by certain bacteria that break down beta-lactam antibiotics, including penicillins and cephalosporins, making these drugs ineffective against infections.
3. Antimicrobial Susceptibility Testing (AST): Laboratory testing used to determine the sensitivity or resistance of bacteria to specific antibiotics. AST results help guide the selection of effective treatments for bacterial infections.
4. Electronic Health Records (EHRs): Digital versions of patients' medical histories maintained over time, including clinical data, test results, and treatments. In this study, EHRs provide patient demographics, clinical and microbiological data.
5. Regional Referral Hospitals (RRH): Tertiary healthcare facilities that provide specialized services to a specific geographic area. These hospitals serve as key data sources in this study for tracking antibiotic resistance patterns.
6. Antibiotic Stewardship: Coordinated interventions and programs aimed at optimizing the use of antibiotics to improve patient outcomes and reduce AMR. Effective stewardship programs promote appropriate antibiotic selection, dosing, and duration.
7. Artificial Intelligence: These are the broader concept of creating machines or systems that can perform tasks that typically require human intelligence, such as problem-solving, decision-making, language understanding, and more.
8. Machine Learning: A subset of AI that focuses specifically on developing algorithms and models that allow machines to learn from data. Instead of being explicitly programmed, ML systems improve their performance over time by recognizing patterns and making predictions or decisions based on the data they process.
9. Microbiological Data: Data derived from laboratory tests that analyze microorganisms (e.g., bacteria, viruses, fungi) and their characteristics.
10. Clinical Data: Data related to the patient's health status, medical history, and treatment.

ABSTRACT

Introduction

In low- and middle-income countries like Uganda, there is growing reliance on empirical prescription of broad-spectrum antibiotics which, while targeting a wide range of pathogens, contributes to the development of resistance to common pathogens such as E. coli. This challenge is compounded by the poor selection of antibiotic panels in many laboratories, which often fail to reflect local resistance patterns and patient-specific factors, leading to inefficient use of scarce resources and delayed appropriate treatment.

Objectives of the study

The objectives of this study were to; 1) identify risk factors for drug resistant E. coli infections using machine learning techniques; 2) evaluate the performance of different machine learning models in predicting the likelihood of drug resistance among patients with E. coli infections using demographic, clinical and microbiological data; and 3) develop a web-based interface to support proper antibiotic prescription and targeted antimicrobial decision-making.

Methodology

A retrospective analysis was conducted on 1,552 records of patients diagnosed with E. coli infections in 10 tertiary healthcare facilities in Uganda. These records were analyzed using machine learning models including Lightgbm, xgboost, random forest, gradient boosting, and decision trees. Feature selection was guided by a weighted importance score and frequency count framework. The best-performing model was deployed in a streamlit-based web interface.

Results

Key predictors of resistance included antibiotic type, patient age, hospital site, specimen type, prior antibiotic use, and hospitalization history. XGboost emerged as the top-performing models for prediction of drug resistance, with an accuracy of 82.32%, a precision of 82.36%, recall of 85.37%, F1 score of 83.84%, and ROC AUC of 90.17%. The web-based interface was implemented using python streamlit technology, intergrated with the best performing model to enable real-time resistance prediction.

Conclusion

This study demonstrates the potential of machine learning to transform antimicrobial resistance surveillance and clinical decision-making in resource-limited settings.

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Introduction

The last century has witnessed a significant increase in life expectancy, largely due to the introduction of effective antimicrobial treatments for infectious diseases (Rosini et al., 2020). However, the emergence and rapid spread of drug resistant pathogens in recent years has complicated the management of these infections, posing one of the most pressing global health challenges today, Antimicrobial Resistance (AMR)(MacIntyre & Bui, 2017, Ayukekbong et al., 2017). AMR occurs when microorganisms such as bacteria, viruses, fungi, or protozoa evolve to resist the effects of antimicrobial agents, rendering previously effective treatments ineffective (Tang et al., 2023). Without intervention, AMR could lead to up to 10 million deaths annually and US$ 1 trillion additional healthcare costs globally by 2050 (Naghavi et al., 2024).

In Low- and Middle-Income Countries (LMICs), the inability to rapidly identify a specific pathogen at the point of care worsens the problem. (McEwen & Collignon, 2018). This diagnostic gap often results into empirical use of broad-spectrum antibiotics which, while targeting a wide range of pathogens, inadvertently contributes to the development of AMR (Mayito et al., 2024; Panda, 2025). Among the most concerning pathogens is Escherichia coli (E. coli), a common gram-negative bacterium, naturally residing in the lower intestinal tract of warm-blooded animals, including humans (Pokharel et al., 2023). E. coli is a major cause of diarrheal diseases, urinary tract infections (UTIs), bloodstream infections leading to sepsis and other invasive conditions such as neonatal meningitis, intra-abdominal infections, and wound infections (Feng et al., 2022). Of particular concern is the emergence of drug-resistant E. coli strains, which exhibit resistance to one or more antibiotics, limiting treatment options and increasing the risk of persistent infections, therapeutic failure, and further complications.

In Uganda, drug resistant E.coli is driven by inappropriate antibiotic use mainly due to over-the-counter dispensing and empirical prescribing without laboratory confirmation (Jackson et al., 2013). This is comfirmed by research that shows 41% of the antibiotic dispensed over the counter in Uganda are issued without a prescription (Bonniface et al., 2021). Even when prescriptions are made in outpatient settings, they are often provided without laboratory confirmation, and empirical use of broad-spectrum antibiotics is common due to delayed antimicrobial susceptibility testing (AST) results (Okello et al., 2020).

This study addresses these challenges by proposing a machine learning based predictive model that leverages demographic, clinical, and microbiological data to predict probability of E.coli drug

resistance in a patient sample. This tool could guide clinicians in selecting appropriate antibiotics during the waiting period, recommend targeted ASTs that prioritize the most promising antibiotics and ultimately improve patient outcomes and antibiotic stewardship in resource-constrained environments.

1.2 Background

Globally, drug resistant E. coli infections account for a substantial proportion of antimicrobial resistance (AMR)-related morbidity and mortality(Daneman et al., 2023). Recent estimates from a global collaboration, which analysed data from systematic literature reviews, healthcare systems and surveillance programs, indicate that in 2019 there was 4.95 million deaths associated with AMR including 1.27 million deaths directly attributed to AMR (Daneman et al., 2023). Among these, E.coli was the leading pathogen, responsible for approximately 829,000 AMR associated deaths and 219,000 AMR attributable deaths (Daneman et al., 2023). The problem is particularly severe in LMICs where drug resistant E.coli shows high prevalence (Mayito et al., 2024; Nkansa-Gyamfi et al., 2019). Drug resistant E.coli infections are associated with increased 30-day and all-cause mortality compared to susceptible infections(MacKinnon et al., 2020). This is intensified by the diagnosis and treatment challenges in this context such as limited access to rapid diagnostic tools and a shortage of second and third-line antibiotics (MacKinnon et al., 2020). The widespread misuse of antibiotics, whether through inappropriate prescriptions, unregulated use in agriculture, or incomplete adherence to treatment regimens, has accelerated the development and spread of drug resistant E.coli in this region (Samy et al., 2022).

Research consistently identifies several key risk factors for drug-resistant E.coli infections. Prior antibiotic use emerges as the most significant predictor, with odds ratios ranging from 1.51 to 21.4 across different populations (Hu et al., 2020; Larramendy et al., 2020; Mitrani-Gold et al., 2023). Previous hospitalization substantially increases risk for hospital-acquired infections (Larramendy et al., 2020). In a study done across all US census regions, age was identified as an independent risk factor for patients having E. coli isolates (Mitrani-Gold et al., 2023). Geographical location in form of travel to high-risk regions significantly increases fecal carriage of drug-resistant E. coli (Hu et al., 2020).

Current approaches to managing these infections rely heavily on rapid diagnostic tools which are often unavailable in health care centers in LMICs (Feng et al., 2022). In this setting, the commonly used antimicrobial susceptibility tests (AST) may take several hours or sometimes days to have test results available (Feng et al., 2022; Segawa et al., 2020). This forces clinicians to resort to using broad-spectrum antibiotics for patient management, which are designed to target a wide array of gram-negative and gram-positive bacteria (Kapisi et al., 2023; Okello et al., 2020). Although these strategies aim to

address the potential susceptibility of multiple pathogens and provide timely treatment to patients, it often leads to the unintended consequences of drug resistance and increased risk of treatment failure.

In response to this growing threat, the country implemented a national AMR surveillance plan aligned with the Global Antimicrobial Surveillance System (GLASS). As part of this initiative, AMR laboratories were established at regional referral hospitals (RRHs) to expand access to antimicrobial susceptibility testing (AST) and strengthen national surveillance capacity (Nabadda et al., 2021). The primary purpose of these laboratories was to enable the execution of AST. However, RRHs in Uganda face unique challenges. As referral centers for large populations across multiple districts, they often experience significant delays, sometimes taking up to several days to receive AST results (Nabadda et al., 2021). During this waiting period, clinicians are forced to rely on empirical treatment with broad-spectrum antibiotics, which increases the risk of further resistance development.

With the adoption of Electronic Health Records (EHRs) and laboratory surveillance systems in Uganda, vast amounts of data including demographic data, clinical data and microbiological data are being collected at the points of care (Nadimpalli et al., 2018). However, this data continues to be underutilized due to challenges such as lack of integrated data systems, limited technical capacity for advanced analytics, and absence of decision-support tools that translate raw data into actionable insights (Hope et al., 2024; Kiggundu et al., 2023; Nsubuga et al., 2024). This gap limits optimized antibiotic prescribing and highlights the need for machine learning to predict E. coli resistance before AST results are available (Nsubuga et al., 2024).

Machine learning (ML) offers a promising approach to analyzing this data, identifying patterns, and reliably predicting probability of drug resistance in patients with E.coli infections (Sakagianni et al., 2023). By leveraging existing EHR and laboratory data at Regional Referral Hospitals (RRH) and tertiary institutions, ML models such as random forests, gradient boosting classifier, extreme gradient boosting (xgboost) and light gradient boosting machine (Lightgbm) can generate actionable insights to inform treatment regimens, support clinical decision-making, and influence policy interventions in Uganda (Gurung et al., 2024; Nedungadi et al., 2024). By integrating machine learning into clinical practice, healthcare providers can identify high-risk patients early, tailor treatments to specific resistance profiles, and reduce reliance on empirical therapy.

This approach aligns with several Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being) and SDG 9 (Industry, Innovation, and Infrastructure) (Gurung et al., 2024; Nedungadi et al., 2024). This alignment highlights the broader impact of integrating AI-driven tools into healthcare contributing to global efforts toward sustainable health and innovation.

CHAPTER 2: LITERATURE REVIEW

This literature review explores the current knowledge on drug resistant E. coli, focusing on its prevalence in Uganda and other low- and middle-income countries (LMICs). It also examines the economic burden of drug-resistant infections, key risk factors contributing to resistance, and the potential of machine learning in predicting these infections. By synthesizing existing research, this section aims to provide a foundation for developing predictive models and context-specific strategies to combat drug resistant E. coli in resource-limited settings.

2.1 Prevalence and Impact of Drug Resistant E. Coli

Understanding the prevalence of drug-resistant Escherichia coli (E. coli) provides a foundation for addressing the problem and proposing effective solutions. Although documentation of drug-resistant E. coli in Uganda remains limited, studies from similar low- and middle-income countries (LMICs) offer valuable insights into its growing burden.

Globally, the rate of drug-resistant E. coli infections is rising faster than the development of new antimicrobials, leading to a shortage of effective treatment options (Martens & Demain, 2017) . This trend is particularly concerning in resource-limited settings, where treatment options are already constrained (M. M. Walker et al., 2022). Several studies have highlighted the high prevalence of drug-resistant E.coli. In Central India, 94% of 710 women attending antenatal clinics carried resistant strains, with multidrug resistance linked to factors such as education level and recent antibiotic use (Pathak et al., 2013). Similarly, in Nigeria, 88.67% of E. coli isolates from pregnant women showed resistance to cephalosporins (Bello et al., 2021). In Ethiopia, a meta-analysis reported an overall E. coli resistance rate of 45.38%, with the highest resistance observed against ampicillin (83.81%) and amoxicillin (75.79%) (Tuem et al., 2018). A multi-country study across Africa and South Asia found that 65% of E. coli isolates from children were resistant to three or more drug classes, with resistance patterns varying by geographic location and antimicrobial usage (Ingle et al., 2018).

Some studies have documented the prevalence of drug resistant E. coli in Uganda, emphasizing its growing burden. A study on carbapenem resistance profiles of pathogenic E. coli in Uganda, which examined 421 isolates, found that 100% of them were drug resistant (Ssekatawa et al., 2021). Another study identified E. coli as one of the most common pathogens in the gastrointestinal tracts of patients visiting outpatient clinics in Kampala and two rural districts. High resistance rates were observed for commonly used antibiotics such as ampicillin and cotrimoxazole (Najjuka et al., 2016). A cross-sectional study at Mbarara Regional Referral Hospital (MRRH), conducted from September 2013 to

June 2014, investigated drug resistant bacterial strains in hospitalized patients. Out of 658 bacterial isolates, 183 (27.8%) were classified as drug resistant, with E. coli accounting for 54.1% of these cases (Ampaire et al., 2015). These findings highlight the significant burden of drug resistant E. coli in health care centers in LMICs such as Uganda and emphasizes how these healthcare settings in act as amplifiers for drug resistant E. coli, necessitating data driven interventions.

2.2. Economic Burden and Need for Early Detection

The economic impact of treating drug resistant infections is substantial. Patients with drug resistant infections face significantly higher hospitalization costs compared to those with non-resistant infections. This is due to the long hospital stays and the heightened cost of second- and third-line medications (Hernandez-Pastor et al., 2023).

A recent study conducted by the Infectious Diseases Institute across nine regional referral hospitals and one tertiary institution, the same sites from which data for this research was obtained revealed that Uganda loses approximately UGX 64 billion (USD 17.7 million) annually due to the wide-ranging costs associated with antimicrobial resistance (AMR). Of this, over USD 12 million is attributed to direct health system costs, including prolonged hospital stays, treatment expenses, and personnel time. Additional losses include USD 4 million in informal care costs and USD 1.4 million in productivity losses due to prolonged illness and premature deaths (Nuwamanya et al., 2025).

Despite the substantial economic burden of drug-resistant infections, valuable data collected at regional referral hospitals remains largely underutilized. This lack of data-driven insights limits the ability of policymakers to quantify the true cost of AMR and evaluate the effectiveness of interventions. Without actionable evidence, health systems struggle to justify investments in advanced solutions such as machine learning tools for early detection and targeted surveillance.

A recommendation from the Mbarara RRH study emphasized the need for preventative strategies, such as improved hand hygiene, use of machine learning tools for early detection and targeted surveillance, to curb the spread of drug resistant E. coli (Ampaire et al., 2015). Given these findings, utilizing machine learning to predict drug resistant E. coli based on clinical, demographic and microbiological information presents an opportunity for early detection and better management.

2.3 Risk Factors for Drug Resistant E.coli

The rise in drug resistance in E. coli is driven by several factors, including improper use of antimicrobial drugs, incomplete adherence to prescribed treatments, and the widespread use of broad-spectrum antibiotics in healthcare and agriculture (Tang et al., 2023). Studies show antibiotic dispensing without prescription is a widespread issue in sub-saharan Africa, with rates as high as 100% in some countries (Sono et al., 2023). Studies on antibiotic prescription practices in Mbarara district highlight this challenge, where clinicians frequently prescribe antibiotics without confirmed diagnostic evidence, worsening resistance trend (Nkansa-Gyamfi et al., 2019; Okello et al., 2020). A meta-analysis found that 69% of antibiotic requests at community drug retail outlets resulted in non-prescription dispensing (Belachew et al., 2021). Patient pressure and failing health systems influence prescription practices, while self-medication is prevalent (Eibs et al., 2020).

In Uganda, one of the primary drivers of antimicrobial resistance is the unregulated, over-the-counter sale of antibiotics without a doctor's prescription, leading to widespread misuse (Ayukekbong et al., 2017). Key drivers include high healthcare costs, convenience, patient demands, and weak enforcement(Sono et al., 2023). Studies have also shown that many drugs dispensed in the region are of questionable quality, further accelerating the development of antibiotic resistance (Basco, 2004). Due to financial constraints, many individuals in low-resource settings seek treatment from traditional healers who provide herbal mixtures of uncertain effectiveness for bacterial infections (“Self-Medication in Rural Africa: The Nigerian Experience,” 2012). Additionally, poor adherence to prescribed treatments is common, with patients often sharing antibiotics, discontinuing medication prematurely, or supplementing prescribed drugs with herbal mixtures believed to enhance efficacy (Misau et al., 2023).

The limited laboratory capacity for antimicrobial susceptibility testing (AST) in many health centers forces clinicians to prescribe antibiotics empirically, often before receiving test results, out of concern for patient mortality. This practice, while intended to provide immediate care, carries significant risks patients may be placed on antibiotics to which the infecting organism is resistant, leading to treatment failure, prolonged illness, and increased risk of complications (Denny et al., 2016). Delays in laboratory testing not only compromise individual patient outcomes but also contribute to the broader spread of resistant infections, as ineffective treatment allows resistant strains to persist and potentially transmit within healthcare settings and communities. In such contexts, the lack of timely diagnostic support undermines antibiotic stewardship efforts and exacerbates the burden of antimicrobial resistance.

Beyond human healthcare, antimicrobials are extensively used in agriculture, particularly in livestock farming, where they are employed for growth promotion, disease prevention and treatment. The overuse of antibiotics in animal husbandry, or as additives in farming, such as in mulching or spraying crops. also contributes to the emergence of drug resistant strains, which can be transmitted to humans through

the food chain, raising significant public health concerns (Naghavi et al., 2024; Richter et al., 2021). In Nepal, a study by Ansari et al revealed that prolonged exposure to low doses of potent antibiotics is a significant risk factor for resistance (Ansari et al., 2015). With limited awareness about antimicrobial resistance in the region, this crisis continues to escalate, requiring urgent intervention.

2.4. Machine Learning for Prediction of drug Resistance.

In high income countries, rapid diagnostic tests (RDTs) have emerged as valuable tools in combating antimicrobial resistance (AMR) across various regions. These tests enable quick and accurate identification of drug-resistant pathogens, facilitating timely treatment decisions and improved clinical outcomes. Despite their potential, LMICs often lack the financial resources to implement such rapid testing on a large scale. (Hur et al., 2024; Nair et al., 2016; Tseng et al., 2017). As a result, drug resistance surveillance in this region remains limited, with delayed diagnoses and increased risk of spread of drug resistant pathogens.

To address this gap, Uganda developed the National Action Plan (NAP) on Antimicrobial Resistance (2024/25–2028/29), aligned with the WHO Global Action Plan on AMR. The NAP outlines strategic interventions to combat drug-resistant infections through a One Health approach, which integrates human, animal, and environmental health. Among its key strategies is the establishment of sentinel surveillance sites, AMR laboratories at tertiary institutions to expand access to antimicrobial susceptibility testing (AST). These laboratories are also tasked with collecting essential resistance data to inform national policy and guide clinical decision-making. However, the data collected through Uganda's AMR surveillance systems is rarely analyzed to identify resistance patterns, potential hotspots, or emerging trends that could inform early interventions, largely due to challenges such as poor data quality, limited digitalization of tools, and inadequate financial and human resources (Kiggundu et al., 2023; Mayito et al., 2024). This underutilization limits the potential of surveillance efforts to guide timely and targeted responses. Compounding this issue is the delay associated with traditional antimicrobial susceptibility testing (AST), which typically requires 24–48 hours to yield results. These delays force clinicians to initiate empirical antibiotic treatment without confirmed resistance profiles, increasing the likelihood of ineffective therapy and accelerating the spread of resistant strains within healthcare settings and communities.

In the past, machine learning has shown great potential in high income countries, predicting drug resistance by leveraging clinical, demographic, and microbiological data (Kherabi et al., 2024; Valavarasu et al., 2025). Machine learning models are computational algorithms that learn patterns from data to make predictions or decisions without being explicitly programmed. These models are trained

on datasets containing features (e.g., demographic data, medical history, clinical findings, and laboratory results) and corresponding labels (e.g., resistance profiles) (Singh* & Mukherjee, 2022). The training process involves optimizing the model's parameters to minimize prediction errors. Once trained, the model can be evaluated on unseen data to assess its performance using metrics such as accuracy (the proportion of all patient cases (both resistant and non-resistant) that the model correctly predicts), precision (among all patients predicted to have drug-resistant E. coli, how many actually have resistance), recall (among all patients who truly have drug-resistant E. coli, how many did the model correctly identify), F1 score (the balance between precision and recall for predicting resistance) and area under the Receiver Operating Characteristic curve (AUC-ROC) (measures how well the model distinguishes between resistant and non-resistant infections across all decision thresholds) (Raj, 2019).

Various machine learning algorithms have been successfully applied to predict antimicrobial resistance (AMR), including ensemble methods such as random forests, gradient boosting, xgboost, and Lightgbm. random forests is an ensemble learning technique that builds multiple decision trees and aggregates their outputs to improve prediction accuracy and reduce overfitting. It is particularly effective in handling high-dimensional data and capturing complex interactions between features, making it suitable for modelling resistance patterns. (Talekar, 2020). gradient boosting builds models sequentially, where each new model corrects the errors of the previous one. This approach enhances performance by focusing on difficult-to-predict cases, which is valuable in datasets with subtle resistance signals (Varghese & K.J., 2024). xgboost (Extreme gradient boosting) is an optimized implementation of gradient boosting that includes regularization techniques to prevent overfitting and improve generalization. It is known for its speed and accuracy, especially in structured data tasks like AMR prediction (Shrivastava et al., 2024). Lightgbm, another gradient boosting framework, is designed for efficiency and scalability (Ke et al., 2017).

ML models implemented in the past were trained on labeled data so that the algorithms learns to map input features (predictors) to the target variable (e.g., resistance status). The dataset is typically split into training and validation sets to ensure the model generalizes well to unseen data (Vermeulen, 2020). Hyperparameters (e.g., learning rate, number of trees in a random forest) are optimized using techniques like grid search or random search to improve model performance (Bergstra et al., 2015). Integrating ML-based prediction into hospital information systems can enable real-time decision support, assisting clinicians in selecting appropriate antibiotics while minimizing unnecessary broad-spectrum use. This approach is particularly helpful in resource-limited settings, where laboratory infrastructure is often inadequate for routine AST. By harnessing ML for drug resistant E. coli prediction, healthcare systems can improve patient outcomes, reduce healthcare costs, and enhance antimicrobial stewardship programs (Hur et al., 2024; Tseng et al., 2017).

While AST remains the gold standard for profiling antimicrobial resistance (AMR) in Uganda's clinical settings, machine learning models have demonstrated the potential to complement these traditional methods by providing highly reliable resistance predictions in significantly shorter turnaround times (Babirye et al., 2024). These predictive tools can guide initial treatment decisions, prioritize samples for AST, and optimize the use of limited laboratory resources thereby enhancing the overall efficiency and impact of AMR surveillance and stewardship efforts . A scoping review on the use of machine learning algorithms for AMR prediction suggests that effective models require inputs such as demographic data, medical history, clinical findings, laboratory AST results. (Moran et al., 2020; Sakagianni et al., 2023). These datasets are processed using machine learning techniques to predict antimicrobial resistance and provide guidance on empirical antibiotic therapy. By leveraging machine learning for AMR prediction, Uganda could enhance antimicrobial stewardship, optimize treatment strategies, and reduce the burden of drug-resistant infections despite existing resource limitations.

2.5 Conclusion

The reviewed literature highlights the growing burden of drug-resistant E.coli infections globally and in Uganda, with significant clinical and economic implications. While studies have documented prevalence, risk factors, and the economic impact of antimicrobial resistance (AMR), several gaps remain. Most existing research in Uganda focuses on descriptive epidemiology and prevalence estimates, with limited emphasis on predictive analytics or early detection strategies.(Hope et al., 2024; Kiggundu et al., 2023). Despite the establishment of AMR surveillance systems and laboratories, data utilization for actionable insights remains minimal due to challenges in digitalization, resource constraints, and analytical capacity (Kiggundu et al., 2023; Mayito et al., 2024; Nabadda et al., 2021). Although machine learning has demonstrated success in predicting resistance in high-income settings, its application in low-resource contexts like Uganda is scarce, particularly using locally relevant data that includes both GLASS and non-GLASS specimen types (Babirye et al., 2024; Nsubuga et al., 2024). These gaps highlight the need for context-specific predictive models that leverage routine demographic, clinical, and microbiological data to support timely decision-making and strengthen antimicrobial stewardship in Uganda.

CHAPTER 3: STATEMENT OF THE PROBLEM, JUSTIFICATION, CONCEPTUAL FRAMEWORK

3.1 Statement of the Problem

In Uganda's tertiary healthcare settings, the diagnosis of drug-resistant E. coli is hindered by the delayed turnaround time of traditional antimicrobial susceptibility testing (AST), often taking several days (Segawa et al., 2020). During this waiting period, clinicians are compelled to initiate empirical antibiotic treatment, which may be ineffective if the patient harbours resistant strains leading to deterioration and increased risk of complications (Zhu et al., 2021). Ivan Segawa et al. (2020) found that in Ugandan health centers, AST was requested in only 4% of cases and performed in just 2.1% of patient files, with an AST turnaround time of 5 days (Segawa et al., 2020). A machine learning model that predicts the likelihood of drug resistance in patients with E. coli infections could help clinicians select antibiotics with a high probability of effectiveness while awaiting AST results.

Although, ASTis guided by national protocols that specify which reagents and antibiotic discs should be used for samples with E.coli growth (Andretta et al., 2024). While these guidelines provide a standardized approach, the core challenge lies in the poor selection of antibiotic panels for testing (Jorgensen, 1993). These panels often fail to reflect local resistance patterns, specimen variability, and patient profiles, resulting in the use of scarce laboratory reagents and antibiotic discs on ineffective options. This mismatch often leads to the use of scarce laboratory reagents and antibiotic discs on ineffective options, causing resource wastage and delaying appropriate treatment (Stalteri Mastrangelo et al., 2021). A data-driven tool that identifies alternative antibiotics with a high probability of effectiveness could assist laboratory technicians in prioritizing these options during AST, thereby conserving limited resources.

Existing studies on use of machine learning tools to predict drug resistant E. coli have primarily used similar data in high-income countries, where it has been successful in guiding priority testing and targeted treatment (Kherabi et al., 2024; Valavarasu et al., 2025). However, these studies have largely focused on GLASS priority specimen types (e.g., urine, blood, stool, urogenital, and cerebrospinal fluid), overlooking the non-GLASS priority specimen types such as pus and tracheal aspirates. The studies were done on data collected from adult populations (over 18 years), leaving significant gaps in representation of the Ugandan context. Uganda being a low income country with over 50% of her

population under 18 years old, these existing studies do not adequately reflect the realities of drug resistant E. coli in Uganda (Uganda Bureau of Statistics, 2024; R. A. Walker et al., 2019).

This research therefore aims to address these gaps by developing a predictive model for drug resistant E. coli infections in persons aged 0 years - 100 years, using demographic, clinical and microbiological data collected in 9 RRHs and 1 tertiary health care institution in Uganda. By incorporating both GLASS and non-GLASS specimens, the project seeks to provide a more comprehensive understanding of drug resistant E. coli patterns. The findings will help identify priority groups and high-risk areas within Uganda's tertiary healthcare institutions, guiding targeted interventions that can reduce drug resistance-related mortality and improve overall patient outcomes.

3.2 Justification

In the past, traditional methods such as clinical assessments and epidemiological surveys were used to identify risk factors for drug resistant E. coli infections. While these approaches provided valuable insights, they lacked the accuracy and context specificity needed for Uganda, which has a distinct population dynamic, healthcare infrastructure, and disease burden. This study moves beyond traditional approaches by applying machine learning (ML) to analyze routine surveillance data from Uganda's tertiary institutions. These ML techniques will enable a more precise, data-driven identification of risk factors for drug resistant E. coli, tailored to the unique Ugandan population and healthcare context. These identified risk factors are expected to directly influence healthcare decisions, enabling clinicians to more effectively prioritize patients for targeted testing and treatments, thus improving patient outcomes and minimizing the misuse of broad-spectrum antibiotics.

Furthermore, this predictive tool will provide evidence-based insights that could inform national policies and AMR stewardship programs, promoting more efficient resource allocation and strategic interventions within the Ugandan healthcare system. By improving the accuracy of drug resistant E. coli predictions and enabling timely clinical interventions, this research will have a significant impact on the SDG 3 (Good Health and Well-being) by reducing mortality from infectious diseases and SDG 9 (Industry, Innovation, and Infrastructure) by leveraging innovation in machine learning to enhance healthcare infrastructure and services in Uganda.

In conclusion, this study represents a necessary step in advancing both clinical practice and health policy in Uganda. Incorporating machine learning to identify and predict drug resistant E. coli infections, this research will strengthen AMR surveillance, and contribute to health policy reforms aimed at combating antimicrobial resistance.

CHAPTER 4: QUESTIONS/ STUDY OBJECTIVES

4.1. Research Questions

1. How can machine learning identify key risk factors for drug resistance in E. coli infections using demographic, clinical and microbiological data ?
2. Which machine learning models demonstrates the highest performance in predicting likelihood of drug resistance in patients with E. coli infections based on demographic, clinical and microbialogical data?
3. How can a web-based interface be developed to predict the likelihood of drug resistance in patients with E. coli infections using machine learning techniques on demographic, clinical, and microbiological data?

4.2 General Objective

To enhance the identification of drug resistance patterns among patients with Escherichia coli infections using patient demographic, clinical, and microbiological data.

4.3 Specific objectives

1. To determine risk factors for drug resistance in patients with E-Coli Infections using machine learning methods on demographic, clinical and microbiological data.
2. To evaluate the performance of different machine learning models in predicting the likelihood of drug resistance among patients with E. coli infections based on demographic, clinical and microbiological data.
3. To develop a web-based interface for prediction of likelihood of drug resistance in patients with E. coli infections using machine learning techniques on demographic, clinical and microbiological data.

CHAPTER 5: METHODOLOGY

In this chapter, I outline the detailed methodical approaches used to achieve the study objectives mentioned above. This includes a description of the study setting, the design approach adopted, the target population, and criteria for selecting study participants. I also describe the methods used to collect relevant data, along with the machine learning tools and techniques applied to analyze the data and obtain the study results.

5.1 Study design

This was a retrospective study of routine laboratory-based surveillance data collected between October 2020 and March 2023. This time frame of October 2020 to March 2023 was chosen because it corresponds to the period during which routine AMR surveillance data was consistently collected across all sentinel sites under Uganda's National Action Plan on AMR. This ensured completeness, comparability, and reliability of data for analysis

5.2 Study area

This study used data from ten tertiary healthcare facilities, comprising nine regional referral hospitals and one tertiary health care institution, located within Uganda. These sites included Jinja Regional Referral Hospital, Department of Medical Microbiology Mbarara University of Science and Technology, Mbarara Regional Referral Hospital, Kabale Regional Referral Hospital, Mbale Regional Referral Hospital, Arua Regional Referral Hospital, Lira Regional Referral Hospital, Gulu Regional Referral Hospital, Masaka Regional Referral Hospital and Soroti Regional Referral Hospital. These facilities were chosen as they provided a representative sample of Uganda's regional healthcare landscape.

Figure 1: Map showing the AMR surveillance sites that were included in this evaluation

5.3 Study Population

The study population consisted of patients who visited the study sites and had cultured samples that demonstrated the growth of pathogens during the study period. It includes both male and female patients aged 0 to 100 years. This study analyzed demographic, clinical, and microbiological data associated with their drug susceptibility test results from blood, cerebrospinal fluid (CSF), urine, pus, and tracheal aspirate specimens cultured during the study period.

5.4 Inclusion and exclusion criteria

5.1.1. Inclusion criteria

i. All patients, both male and female, aged 0 years - 100 years whose samples were collected and cultured at any of the study sites.

5.1.2. Exclusion criteria

i. Patients with more than 80% missingness in the required fields of the data.

5.5 Data description

This study analyzed retrospective data collected by qualified clinical and laboratory teams at all surveillance sites. This dataset encompassed both GLASS priority specimens (urine, blood, stool, urogenital, and cerebrospinal fluid [CSF]) and non-GLASS specimens, such as pus and tracheal aspirates (Mayito et al., 2024). The data was accessed through Uganda's AMR Data warehouse (https://amrdb.idi.co.ug/). A formal request was submitted together with the study concept and School of Public Health Research Ethics Committee (REC) approval. Upon review and clearance, the relevant dataset was provided for analysis. This process ensured compliance with ethical and institutional requirements.

The original dataset contained results from 20,063 patients and 84 variables, including 64 antibiotic susceptibility test (AST) results. After filtering for the target pathogen, Escherichia coli, the wide format dataset was reduced to 1,552 patients. This dataset was then reshaped into a long format (a data structure where each row represents a single observation) based on the antibiotics, enabling each row to represent a unique antimicrobial exposure per patient sample. This structure allowed for more granular analysis of antibiotic usage patterns and antimicrobial susceptibility test outcomes (Haredasht et al., 2025). This resulted into a data frame of 10,181 rows containing a unique observation of the sample's interaction with an antibiotic.

The data columns in the resultant dataset captured a wide range of information including patient ID (a unique identifier assigned to each patient), Lab (the health facility where the sample was collected and processed), Specimen number (unique code assigned to each laboratory specimen), Specimen type (the biological material collected for testing), Sex (gender of the patient) , Age (the patient's age in complete years at the time of sample collection) , Department (the hospital ward or unit where the patient was admitted), Hospitalization for more than 48 hours, Collection date (the date on which the specimen was collected), Transferred from another facility (whether the patient was referred from another health facility), Prior Antibiotic Therapy (whether the patient had been exposed to antibiotics before sample collection), Duration on antibiotics (the number of days the patient had been on antibiotics prior to sample collection), Date of Admission (the date the patient was admitted to the hospital), Diagnosis (the clinical diagnosis recorded at the time of admission or sample collection) and antimicrobial susceptibility tests across different antibiotics for each patient. The structure of the resultant dataset which was used for analysis in this study is as shown below.

```
tibble [10,181 × 18] (S3: tbl_df/tbl/data.frame)
 $ ID                          : chr [1:10181] "ARU-000-8-FL" "ARU-000-8-FL" "ARU-000-8-FL" "ARU-00-13-FL" ...
 $ Lab                         : num [1:10181] 0 0 0 0 0 0 0 0 0 0 ...
 $ Organism                    : chr [1:10181] "Escherichia coli" "Escherichia coli" "Escherichia coli"
"Escherichia coli" ...
 $ Sex                         : chr [1:10181] "0" "0" "0" "1" ...
 $ Age                         : num [1:10181] 10 10 10 18 18 18 38 38 38 38 ...
 $ Department                  : chr [1:10181] NA NA NA "Out Patient Department" ...
 $ Collectiondate              : chr [1:10181] "01/12/2020" "01/12/2020" "01/12/2020" "07/10/2020" ...
 $ Specimentype                : num [1:10181] 1 1 1 0 0 0 0 0 0 0 ...
 $ Comment                     : chr [1:10181] NA NA NA NA ...
 $ Hospitalized48hrs           : num [1:10181] 2 2 2 0 0 0 1 1 1 1 ...
 $ Transferredfromanotherfacilit: num [1:10181] 2 2 2 0 0 0 0 0 0 0 ...
 $ Priorantibiotictherapy      : num [1:10181] 2 2 2 0 0 0 1 1 1 1 ...
 $ Durationofantibioticsindays : num [1:10181] 0 0 0 0 0 0 0 0 0 0 ...
 $ AdmissionStatus             : num [1:10181] 0 0 0 0 0 0 0 0 0 0 ...
 $ Diagnosis                   : chr [1:10181] "Skin and Soft Tissue:" "Skin and Soft Tissue:" "Skin and Soft
Tissue:" "Completely Unknown Indication" ...
 $ Antibiotic_Code             : chr [1:10181] "AMC_ND20" "CIP_ND5" "SXT_ND12" "CIP_ND5" ...
 $ Resistance                  : chr [1:10181] "S" "S" "R" "S" ...
 $ AST_Result                  : num [1:10181] 0 0 1 0 0 0 1 1 1 1 ...
[1] 10181
```

Figure 2: Screenshot of the tribble structure for the data used for prediction of drug resistance in patients with E.coli infections

5.7 Data Preprocessing

During data preprocessing, an outcome variable, Antimicrobial Susceptibility Testing (AST) Result was created and dichotomized into resistant and non-resistant categories. Isolates classified as R (resistant, where the antibiotic is unlikely to be effective and alternative treatment is required) were coded as 1, indicating a high probability of treatment failure. Isolates classified as S (susceptible, where the antibiotic is likely to be effective at standard doses) and I (intermediate, where the antibiotic may be effective under specific conditions such as higher doses or targeted delivery) were coded as 0, indicating likelihood of the antibiotic being effective under standard dosing or specific conditions such as higher dosing. This classification was in accordance with internationally recognized standards for antimicrobial susceptibility testing, such as those provided by the European Committee on

Antimicrobial Susceptibility Testing (European Committee on Antimicrobial Susceptibility Testing (European Committee on Antimicrobial Susceptibility Testing (EUCAST), 2025). EUCAST clinical minimum inhibitory concentration breakpoints categorize organisms as susceptible at both S and I, with S indicating susceptibility at standard dosing and I indicating susceptibility at increased exposure (European Committee on Antimicrobial Susceptibility Testing (European Committee on Antimicrobial Susceptibility Testing (EUCAST), 2025). This binary coding facilitated predictive modeling while maintaining clinical relevance and microbiological standards

Upon further inspection of the outcome variable, antimicrobial susceptibility test result, the value entry '6' was identified as an outlier that did not correspond to any recognized antimicrobial susceptibility test classifications (S, I or R). Since it appeared only once in the dataset and lacked a clear interpretation, it was excluded from the analysis to preserve data integrity and transparency.

Handling Missing Data

Missing values were addressed using imputation techniques tailored to the nature and distribution of each variable. For numerical features, mean or median imputation was applied depending on the skewness of the data. For categorical variables, mode imputation was used where appropriate. Variables with excessive missingness were either imputed by prediction using suitable models or excluded from the analysis due to their limited utility and potential to introduce bias.

Feature Selection and Reduction

Several variables were dropped based on redundancy, irrelevance, or data quality concerns:

- Day, Month, and Year were removed as they were already encapsulated within the Collection Date variable.
- Department was dropped due to inconsistent and uncleanable entries, which would have required re-verification with data collectors and hospital records.
- Organism was excluded as it was redundant; the dataset had already been filtered to include only E. coli cases under the Organism variable.

Normalization and Encoding

To ensure that all features contributed proportionately to the model, categorical variables including sex, hospitalization for more than 48 hours, transfer from another facility, antibiotic code, and specimen type, were encoded using Label Encoding, which assigns a unique integer to each category. This method was chosen for its simplicity and compatibility with tree-based models.

5.6 Study variables

5.1.3. Dependent variable/Outcome variable

The outcome variable is the presence of drug resistance in patients with E. coli infections.

5.1.4. Independent variables/Exposures

The independent variables include the patient sex, age in complete years, hospital ward, date of sample collection along with the day and sample type. Additional variables include whether the patient was hospitalized for more than 48 hours, whether they were transferred from another facility, if there was prior antibiotic therapy, and the duration of antibiotic use in days if applicable.

5.8 Feature Engineering

Additional derived features were created to enrich the dataset and capture the antibiotics and outcome variable. Admission status was defined based on the presence of an admission date in the patient record, which reliably indicates that the patient was hospitalized. This variable was retained for its clinical relevance, as hospitalization increases exposure to healthcare-associated infections and resistant organisms. Therefore, using admission status was a justified approach during feature engineering, ensuring that this important risk factor was not discarded simply due to incomplete discharge data. Admission status had values such as yes, no, NA.

5.9 Feature Selection

To address the first objective, determining risk factors for drug resistance in patients with E-Coli infections, a machine learning approach was applied to a long-format dataset derived from microbiological, demographic, and clinical records. Four machine learning models, random forest, gradient boosting, xgboost, and logistic regression were trained on this data to identify features most predictive of resistance, using a binary outcome variable, antimicrobial susceptibility test result. These models were selected to balance interpretability, predictive power, and the ability to capture complex patterns in the data. logistic regression was included for its simplicity and transparency, offering easily interpretable coefficients that help explain the influence of individual predictors (Chui & Chan, 2025).

Using a diverse set of algorithms enabled cross-validation of insights and a comparative analysis of feature importance, helping to identify consistently predictive factors across different modeling approaches. This triangulation strengthened the robustness and generalizability of the findings, ensuring that conclusions were not model-dependent but reflected underlying data patterns (Veledar et al., 2025). Feature selection was guided by the WISFC (Weighted Importance Score and Frequency Count)

framework, a method that aggregates feature importance scores from random forest, gbm, xgboost, and logistic regression, accounting for both the frequency of feature appearance and its average importance (Veledar et al., 2025). The resulting WISFC scores provided a consensus-based ranking that supports more reliable interpretation.

To explore the relationships between the key demographic, clinical, and microbiological variables and the correlation matrix was plotted, showing the strength and direction of relationships between the variables. The color scale ranges from -1 (strong negative correlation) to +1 (strong positive correlation), with darker orange indicating stronger negative relationships and darker blue indicating stronger positive ones. This analysis was performed to identify pairs of features with high linear relationships. Variables exhibiting strong correlations (typically above a threshold such as 0.8) were considered for removal to prevent redundancy and reduce the risk of multicollinearity, which can distort model coefficients and impair interpretability.

5.10 Machine Learning Models

Five machine learning models were evaluated to predict drug resistance, including Decision Trees for rule-based classification, random forests to enhance performance through ensemble learning, gradient boosting for sequential error correction, xgboost for its optimized performance and feature importance capabilities, and Lightgbm booster, which offers efficient training and scalability for large datasets. These models were chosen for their ability to handle complex data structures, enabling a robust comparison of predictive performance (Al Musyaffa et al., 2025; Chung et al., 2023; Halloran, 2009; Pratiwi et al., 2024).

5.11 Model Training and Evaluation

The models were trained using 80% of the dataset, while the remaining 20% will be reserved for testing to evaluate performance. The 80/20 split is a widely accepted standard in machine learning and statistical modelling for training and testing datasets. It ensures that the model has sufficient data to learn patterns while retaining a portion for unbiased evaluation (Gholamy et al., 2018). Model performance was evaluated using multiple metrics, including accuracy (the proportion of correct predictions out of all predictions made), precision (the proportion of true positive predictions among all positive predictions, indicating how many predicted positives were actually correct), recall (the proportion of true positives identified out of all actual positives, showing the model's ability to capture relevant cases), F1-score (the harmonic mean of precision and recall, providing a balanced measure when both metrics are important). Additionally, the receiver operating characteristic (ROC) curve and

area under the curve (AUC) utilized to assess the trade-off between sensitivity and specificity, with the AUC serving as an indicator of overall model performance (Riyanto et al., 2023). A confusion matrix was used to visually inspect the number of false positives, false negatives, true positives, and true negatives (Canbek et al., 2021).

5.12 Model Evaluation and Selection

Each model was evaluated across a defined parameter grid, and the best-performing configuration was selected based on the F1 score. The final models were then assessed on the test set using accuracy, precision, recall, F1 score, and ROC AUC metrics. (Bergstra & Bengio, 2012).

5.13 Design and Implementation of the Web-Based Interface

To facilitate real-time prediction of drug resistance in patients with Escherichia coli infections, a web-based interface was developed using Streamlit and Python, leveraging open-source technologies to ensure flexibility, transparency, and scalability. This approach enables seamless integration with existing health information systems and provides a foundation for incorporating advanced analytics, including artificial intelligence (AI) and machine learning models. The tool is intended for use after pathogen identification and before antimicrobial susceptibility testing (AST), providing clinicians with timely preliminary insights during the waiting period for AST results. The use of Streamlit allowed for rapid prototyping and deployment of an interactive, web-based tool that can be easily adapted to different clinical settings. This aligns with global trends in digital health innovation as reported in literature and supports the development of context-aware clinical decision-support systems (Fabrizzio et al., 2023).

The best performing model was integrated into the interface capable of analyzing patient data and predicting the likelihood of antimicrobial resistance. The design emphasized a clean, intuitive layout for healthcare professionals with minimal technical background. Users can input clinical and demographic variables, and the application returns a prediction along with confidence scores. The interface is modular, allowing for future integration with hospital information systems or national AMR surveillance platforms. The flowchart below illustrates the workflow of the web-based AMR prediction interface.

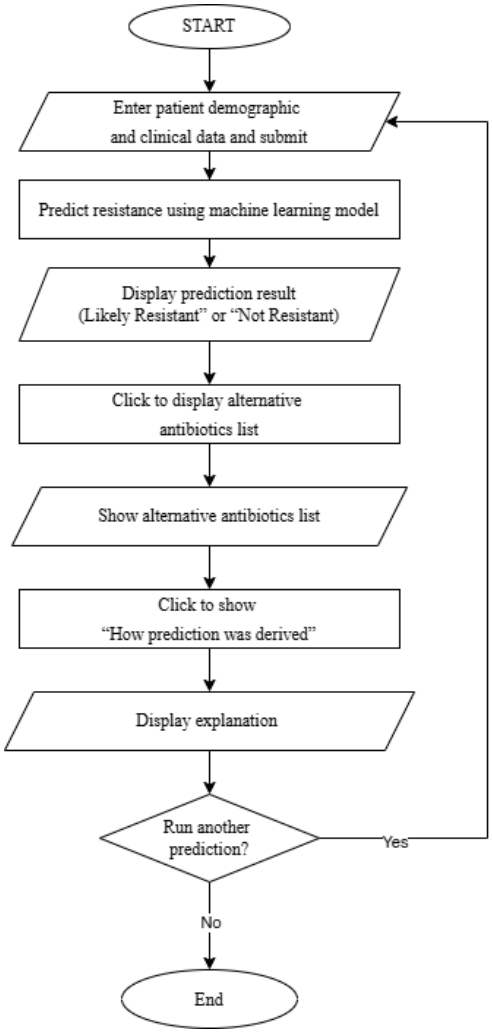

Figure 3: Flowchart of AMR Prediction Workflow from Data Entry to Resistance Prediction and other options

The process begins with the user entering patient demographic and clinical data, such as age, sex, prior antibiotic use, hospitalization history, and specimen type, followed by submission. Once the data is submitted, the system applies the best performing model to predict whether the patient's sample is likely resistant or likely not resistant to antibiotics. The prediction result is then displayed on the interface, providing immediate feedback to the user. To enhance usability and support clinical decision-making, the interface offers additional interactive features. Users can click to view a list of alternative antibiotics that are likely to be effective based on the prediction outcome. Furthermore, the system provides an option to display an explanation of how the prediction was derived, including confidence scores or feature importance, promoting transparency and trust in the model. Finally, the workflow includes a decision point where the user can choose to run another prediction or end the process, ensuring flexibility and iterative use of the tool in clinical settings.

5.14 Ethical consideration

This study adhered to ethical guidelines to ensure patient confidentiality, data security, and compliance with legal and institutional regulations. Patient data was anonymized, securely stored, and accessed only by authorized personnel. The School of Public Health Institutional Review Board (IRB) approval was obtained. Measures were taken to minimize bias in data processing and model training, ensuring fairness in predictions. The developed machine learning model can serve as a clinical decision-support tool rather than a sole determinant in treatment or laboratory testing decisions.

5.15 Conceptual Framework

The conceptual framework illustrates the interaction between contextual factors, demographic, clinical and microbiological factors and the antimicrobial resistance outcome (Resistance or Non-Resistance) in patients with E. coli infections.

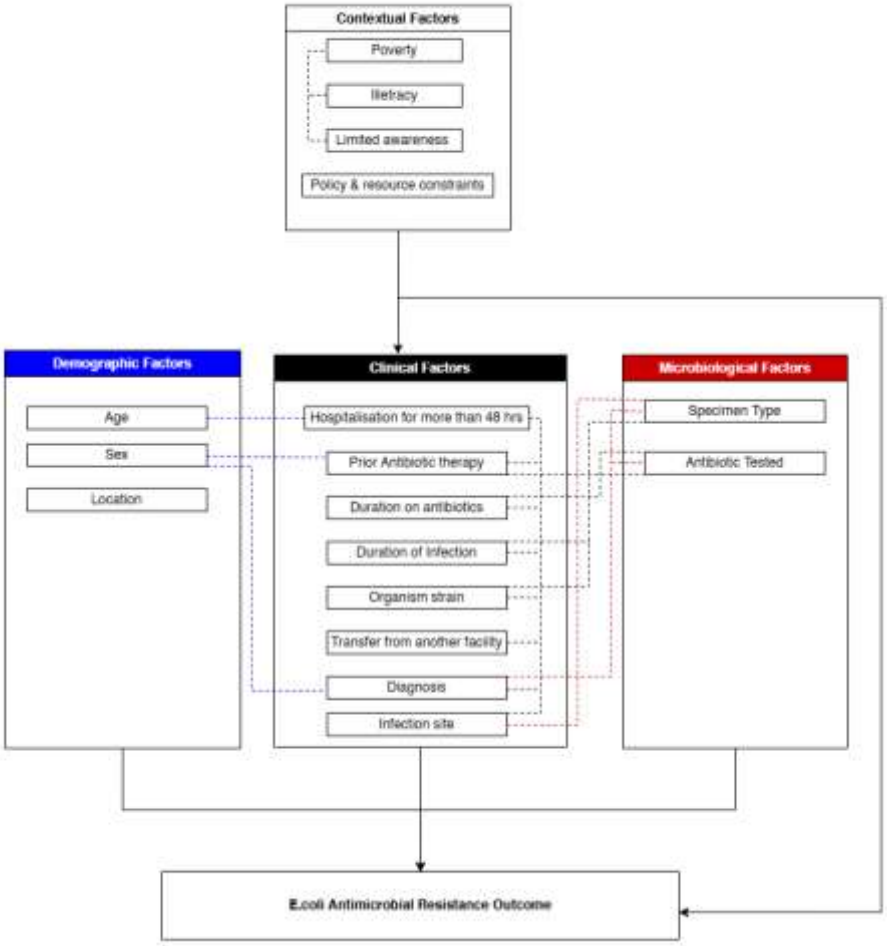

Figure 4: Conceptual Framewok illustrating the interaction of various variables and their influence on drug resistance outcome

Resistance outcome (Resistance or Non -resistance) to an antibiotic is largely dependent on 4 categories of factors, contextual, demographic, clinical and microbiological factors. Contextual factors which represent broader social and systemic influences that indirectly affect antimicrobial resistance include poverty, which limits access to healthcare and promotes self-medication; Illiteracy reduces understanding of proper antibiotic use; limited awareness which leads to misuse and overuse of antibiotics; policy and resource constraints which affect laboratory capacity and timely diagnostics. These factors influence clinical practices and patient-level exposures, creating conditions that increase AMR risk. Demographic factors which are individual characteristics that may predispose patients to resistant infections which include Age, Sex, and Location which provides insight into the age, gender and geographical differences in resistance patterns. Clinical Factors are the variables related to patient history and hospital experience which include Hospitalization for more than 48 hours, Prior antibiotic therapy, Duration on antibiotics, Duration of infection, Organism strain, Transfer from another facility, Diagnosis, Infection site. Microbiological Factors which include laboratory-related variables that influence resistance outcomes like Specimen type (e.g., urine, blood, pus), Antibiotic tested during AST. All these factors converge to predict the E. coli antimicrobial resistance outcome, which is the dependent variable in the study.

5.16. Analytical Framework

The analytical framework outlines the process of developing a machine learning model for predicting antimicrobial resistance in E. coli infections using demographic, clinical, and microbiological data

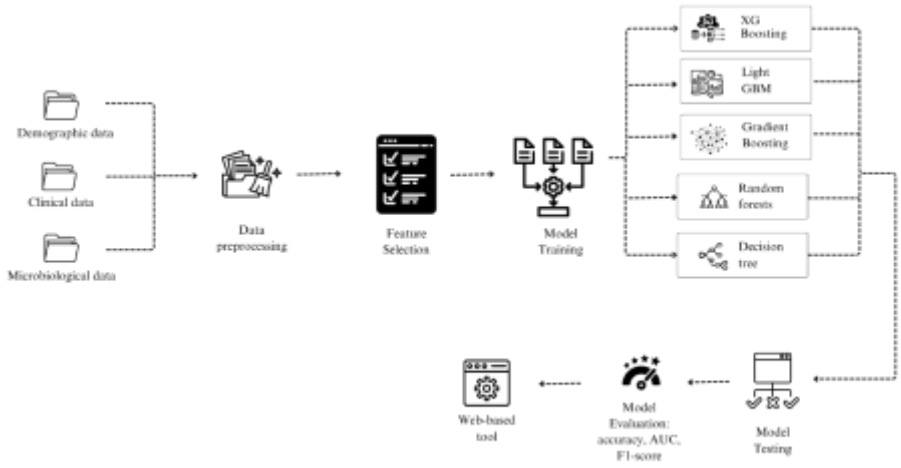

Figure 5: An analytical framework for prediction of drug resistant E.coli

The independent variables include demographic, clinical, and microbiological factors. Demographic factors such as age and sex provide essential patient characteristics that influence drug resistant risk.

Clinical factors such as hospitalization history, prior antibiotic therapy, duration of antibiotic use, and hospital ward, that might contribute to the likelihood of hosting drug resistant pathogens was included. Microbiological factors such as the specimen type, antibiotic being tested and antimicrobial susceptibility test (AST) results will also be included.To ensure high-quality data for analysis, pre-processing step such as handling missing values, encoding categorical variables, normalizing continuous features, and feature selection was applied. This approach optimized the dataset for machine learning models, ensured that the dataset was clean and well structured. This involved handling missing values, encoding categorical variables, normalizing continuous features, and selecting the most relevant predictors using statistical techniques.

Various algorithms, including lightgbm booster classifier, random forest, gradient boosting and xgboost, decision trees were trained and evaluated to predict the drug resistant status based on the predictor variable entries. The best-performing model was selected based metrics such as accuracy, precision, recall, F1-score, and AUC-ROC analysis. The best-performing model was serialized and saved as a file, which is loaded by the Streamlit application during runtime to generate predictions, allowing healthcare professionals to input patient data and receive real-time drug resistance predictions.

CHAPTER 6: RESULTS

In this section, I present the results of this retrospective study on drug-resistant Escherichia coli infections which aimed to evaluate the effectiveness of machine learning in identifying risk factors associated with resistance, developing a predictive machine learning model using demographic, clinical, and microbiological data, and creating a web-based interface for the best-performing model. The findings are organized according to the study objectives and demonstrate the potential of machine learning to improve early detection, guide targeted antimicrobial susceptibility testing, and support clinical decision-making in resource-limited settings.

6.1 Dataset summary

6.1.1 Geographical Distribution of Sample Collection Sites

The observations of patients whose samples were observed to have E. coli growth belonged to 9 Regional Referral Hospitals (RRH) namely; Mbarara Regional Referral Hospital (361), Kabale Regional Referral Hospital (222), Lira Regional Referral Hospital (183), Mbale Regional Referral Hospital (159), Jinja Regional Referral Hospital (129), Gulu Regional Referral Hospital (100), Arua Regional Referral Hospital (86), Masaka Regional Referral Hospital (66), Soroti Regional Referral Hospital (62) and 1 Tertiary Hospital at Department of Medical Microbiology Mbarara University of Science and Technology (DMM MUST) (184).

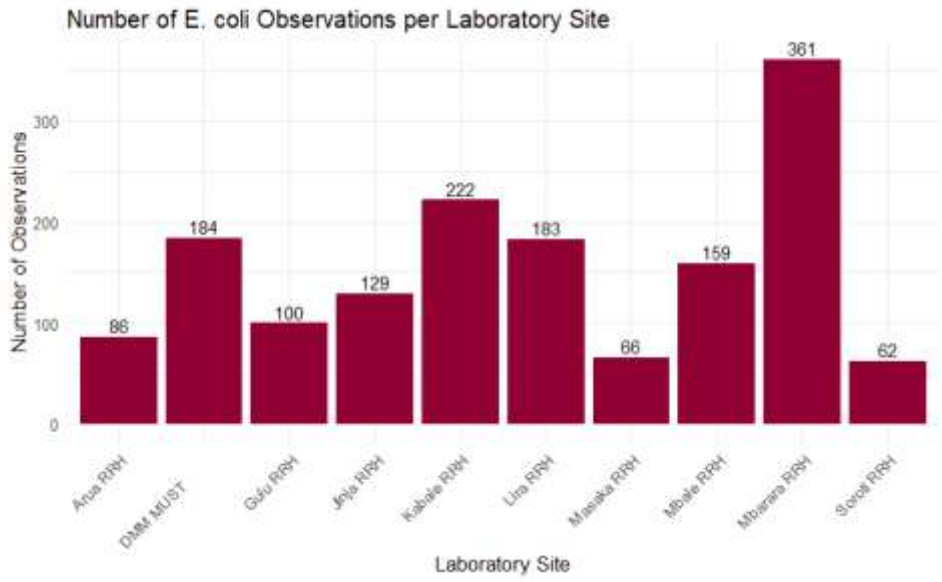

Figure 6: Distribution of Escherichia coli Observations Across Laboratory Sites

6.1.2 Age and Sex Distribution of patients with samples that had E. coli growth

The distribution of client sex in the dataset was imbalanced, with females comprising 67% (n = 1,040) of the total sample. Female clients had a lower average age of 32.9 years (SD = 18.8), compared to male clients whose average age was 42.8 years (SD = 23.3). Among the female clients analyzed, 532 (51%) were aged between 22 and 42 years, while 262 male clients (51.2% of the males analyzed) were between 28 and 60 years. The median age was 30 years for females and 42 years for males. A visual representation of these age distributions is provided in the figure below.

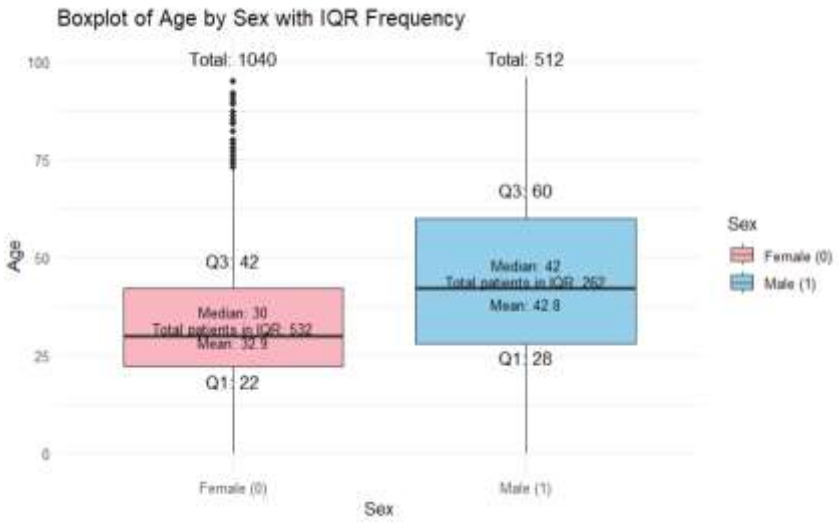

Figure 7: Boxplot showing Age and Sex Distribution of patients with samples that had E.coli growth

Below is a summary table showing the age and sex distribution of patients with samples that had E. coli growth.

Table 1: A summary table showing the age and sex distribution of patients with samples that had E.coli growth

Sex	Mean_Age	Median_Age	Q1	Q3	Count_IQR	Total
Female (0)	32.9	30	22	42	532	1040
Male (1)	42.8	42	28	60	262	512

6.1.3 Distribution of observations across the outcome variable

Out of the total observations (10,181) in the resultant long format dataset, 4709 were categorized as non-resistant (0), while 5471 were classified as resistant (1). This moderate imbalance is illustrated in the bar chart below:

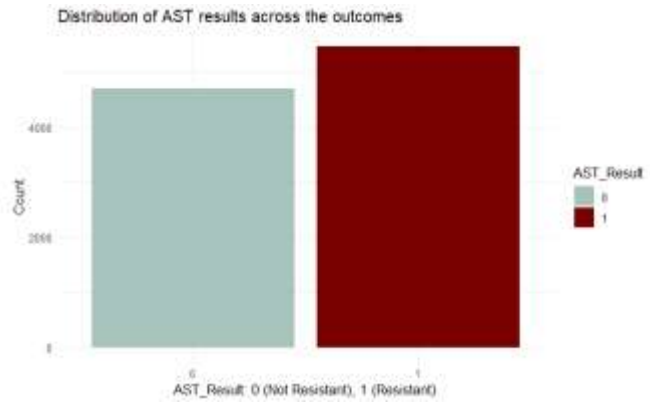

Figure 8: Bar chart showing the distribution of categorical resistance outcomes

Using the 10,180 samples to analyze the association between patient sex and antimicrobial susceptibility test result outcomes, results showed that among female patients (0), 52.4% (3,457 samples) exhibited resistance while 47.6% (3,140 samples) were not resistant. Among male patients (1), 56.2% (2,014 samples) were resistant, and 43.8% (1,569 samples) were susceptible.

```
Contingency Table (Sex vs AST_Result):
            Non-Resistant (0)  Resistant (1)
Female (0)               3140           3457
Male (1)                 1569           2014

Proportions of Resistance by Sex:
            Non-Resistant (0)  Resistant (1)
Female (0)           0.475974       0.524026
Male (1)             0.437901       0.562099
```

Figure 9: Distribution of antimicrobial resistance (AST result) by sex, where Sex = 0 represents females and Sex = 1 represents males. AST result = 1 indicates resistance, and 0 indicates susceptibility.

6.2 Objective 1: Identify the risk factors for drug resistant E. coli infections

6.2.1 Feature importance and selection

Across all models, a unique set of features consistently emerged as important predictors of drug resistance. These include the number of days the patient had been on antibiotics prior to sample collection, a particular antibiotic the patient was on before sample collection, the patient's age in complete years at the time of sample collection, the health facility where the sample was collected and processed, the specimen type collected for testing, whether the patient was hospitalized for more than 48 hours, whether the patient had been exposed to antibiotics before sample collection, the gender of the patient, the patients admission status, and whether the patient was transferred from another health facility.

Tree-based models, random forest, gradient boosting, and xgboost placed greater emphasis on microbiological and demographic variables; duration of a patient on antibiotics, antibiotic being tested, age, the health facility where the sample was collected and processed, and specimen type consistently ranked as the top four predictors. In contrast, the logistic regression prioritized clinical history, identifying prior antibiotic therapy, hospitalisation for more than 48 hours and patient having been transferred from another facility as its most influential features. Figure 8,9,10 and 11 below visualize these differences in feature importance across models.

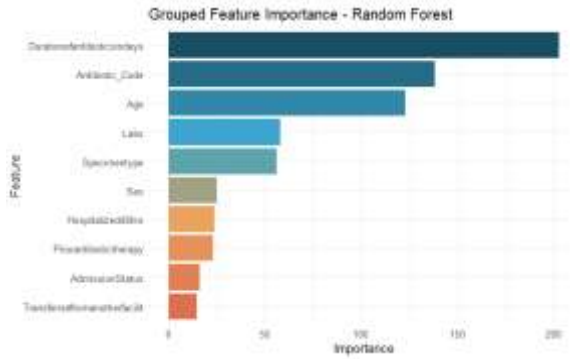

Figure 10: Random forest model showing the top predictors of drug resistance in E-Coli infections

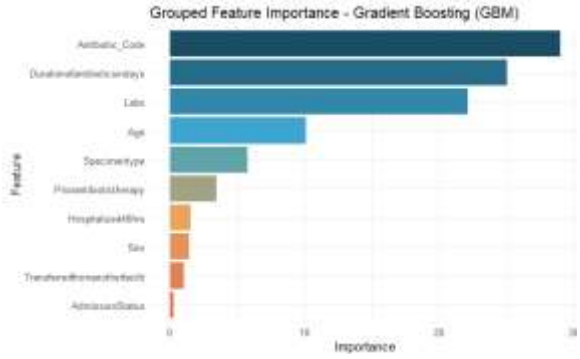

Figure 11: Gradient boosting model showing the top predictors of drug resistance in E-Coli infections

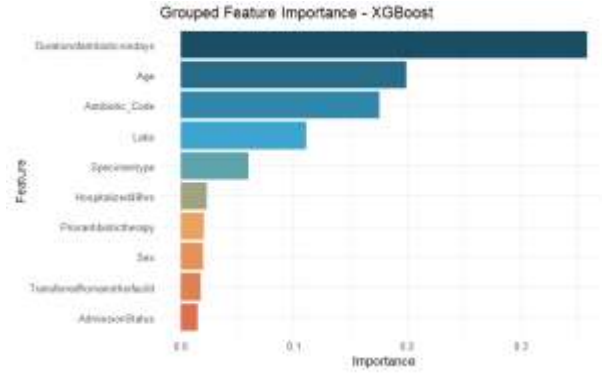

Figure 12: XGBoost model showing the top predictors of drug resistance in E-Coli infections

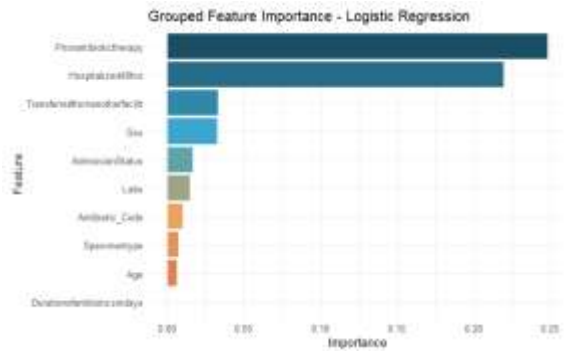

Figure 13: Logistic regression model showing the top predictors of drug resistance in E-Coli infections

6.2.2 Triangulation of the top 7 features using the Weighted Importance Score and Frequency Count framework

As visualized in the figure below, the seven top-ranked features selected for model development were the number of days the patient had been on antibiotics prior to sample collection, a particular antibiotic tested on the sample, the patient's age in complete years at the time of sample collection, the health facility where the sample was collected and processed, the specimen type collected for testing, whether the patient was hospitalized for more than 48 hours, and whether the patient had been exposed to antibiotics before sample collection.

Figure 14: A feature selection across random forest, gbm, xgboost and logistic regression

6.2.3 Exploring Variable Relationships Using a Correlation Matrix

The correlation matrix is as shown in the figure below:

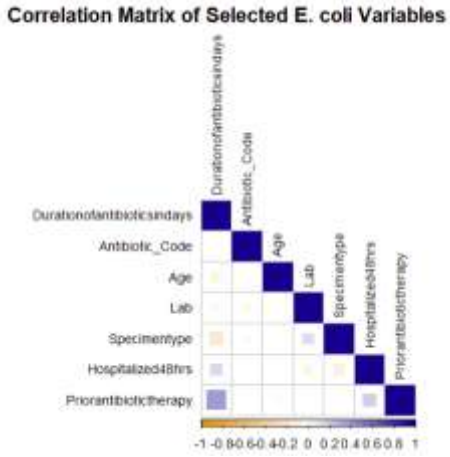

Figure 15:Heatmap of the correlation matrix showing relationships among the top predictors of AMR in patients with E.coli infections.

Most variables had weak correlations (close to 0), suggesting they provide independent information for modelling. The duration of a patient on antibiotics showed a positive correlation with a patient's history of prior antibiotic intake, a moderate positive correlation with a patient's history of hospitalisation for more than 48 hours, and a moderate negative correlation with the specimen type. There was a moderate positive correlation between patient's history of hospitalisation for more than 48 hours and patient's history of prior antibiotic therapy. Overall, the absence of very strong correlations (>0.8) reduces concerns about multicollinearity, meaning most variables can be retained for machine learning models without redundancy.

6.3 Objective 2: To evaluate the performance of different machine learning models in predicting the likelihood of drug resistance among patients with E. coli infections

Ensemble methods consistently outperformed the decision trees across all performance metrics with xgboost as the top performing model. The results obtained for each model are presented in figure below:

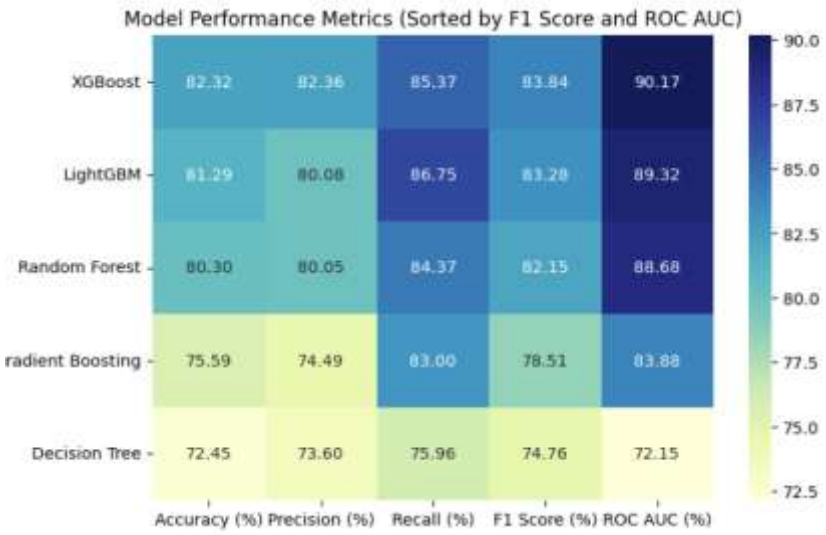

Figure 16:Model performance sorted by F1score and ROC AUC

Confusion Matrix

The confusion matrices revealed differences in predictive performance across the other models. xgboost and Lightgbm outperformed the others, showing the highest number of correctly predicted resistant cases and the lowest wrongly predicted resistant cases. Lightgbm had the highest number of correctly predicted drug resistant cases (949) with the lowest incorrectly predicted drug resistant cases (145). xgboost performed better in correctly identifying non resistant cases while it only correctly identified 934 actual

resistant cases and wrongly classified 160 resistant cases as non-resistant. Below are the confusion matrices for each of the achine learning models evaluated.

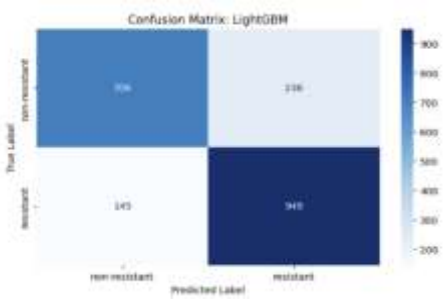

Figure 17: Confusion Matrix for the Lightgbm Classifier model showing classification performance in predicting antibiotic resistance in patients with E.coli

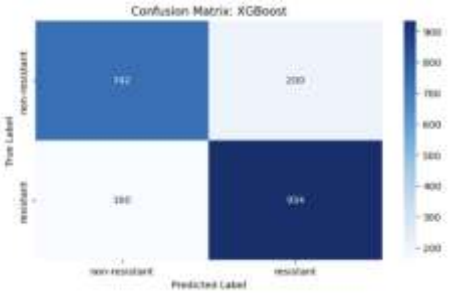

Figure 18: Confusion Matrix for the xgboost model showing classification performance in predicting antibiotic resistance in patients with E.coli

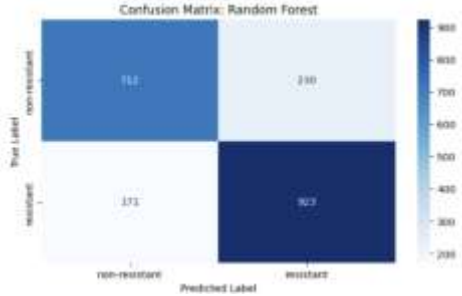

Figure 19: : Confusion Matrix for the random forest model showing classification performance in predicting antibiotic resistance in patients with E.coli

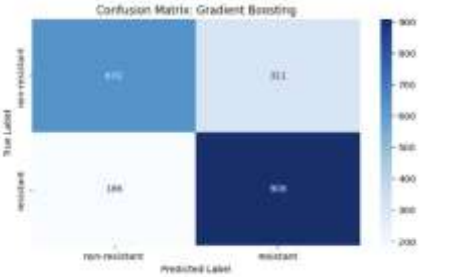

Figure 20: Confusion Matrix for the gradient boosting model showing classification performance in predicting antibiotic resistance in patients with E.coli

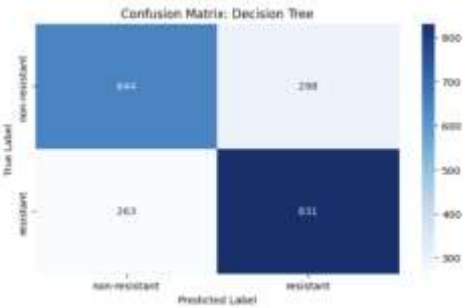

Figure 21: : Confusion Matrix for the decision tree model showing classification performance in predicting antibiotic resistance in patients with E.coli

6.4 Objective 3: To develop a web-based interface for prediction of existence of drug resistance in patients with E. coli infections.

The web-based interface was designed for two primary user groups: medical doctors and laboratory staff. Medical doctors interact with the interface by entering patient demographic and clinical data, such as age, sex, prior antibiotic use, hospitalization history, and specimen type. After submitting the data, they receive a real-time prediction indicating whether the patient's sample is likely resistant or not resistant to the antibiotic entered. Doctors can then view a list of alternative antibiotics with a higher likelihood of effectiveness to support prescribing decisions while awaiting susceptibility test results.

Laboratory staff use the interface to identify antibiotics that the patient is most likely susceptible to, enabling them to prioritize these drugs during antimicrobial susceptibility testing, promoting efficient use of antibiotic discs. The interface includes interactive features such as buttons to display alternative antibiotics and explanations of how predictions were derived. This design ensures usability, transparency, and supports informed decision-making for both clinicians and laboratory personnel. Below are screenshots of the web based tool.

Figure 22: A screenshot of the web-based tool for AMR resistance (Likely resistant to antibiotics)

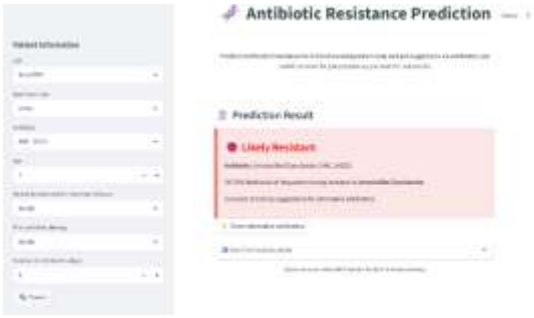

Figure 23: A screenshot of the web-based tool for AMR resistance (Not resistant to antibiotics)

CHAPTER 7: DISCUSSION

This study successfully applied machine learning (ML) techniques to predict drug resistance in Escherichia coli infections using demographic, clinical, and microbiological data from 10 tertiary healthcare facilities in Uganda. The results demonstrate that ML models, particularly ensemble methods like xgboost and Lightgbm, can effectively identify resistance patterns and support clinical decision-making for patients with E. coli infections in regional referral hospitals in Uganda.

7.1 Dataset summary

7.1.1 Geographical Distribution of Sample Collection Sites

The geographical distribution of these sites, as visualized in the results section, reflects a broad national coverage, capturing data from all This diversity in sampling locations enhances the representativeness of the findings and allows for regional comparisons in antimicrobial resistance (AMR) patterns (MacFadden et al., 2016). However, the number of E. coli observations varied significantly across sites. From the public health perspective, this variation could suggest the presence AMR hotspots within certain districts. Yet, it may also reflect disparities in laboratory capacity, patient volume, or diagnostic practices across facilities (Bender et al., 2004; Obakiro et al., 2021). These differences raise concerns about potential underreporting in lower volume sites, which may not necessarily indicate lower AMR prevalence but rather limitations in surveillance infrastructure (Nakate et al., 2022). Research by Kiggundu et al identified existing gaps in AMR surveillance data collected at sentinel sites, some of which were data sources for this study, that could obscure the true extent of burden of AMR (Kiggundu et al., 2023). Such gaps risk leading to misinformed policy or inefficient resource allocation (Kiggundu et al., 2023). Addressing these disparities is important to ensuring that AMR surveillance systems are both equitable and actionable (Kiggundu et al., 2023).

7.1.2 Age and Sex Distribution of Clients

The demographic analysis of the study population revealed a notable sex imbalance, with females accounting for 67% (1,040) of the total sample and males comprising 33% (502). This imbalance may reflect underlying gender dynamics in healthcare-seeking behaviour, with women potentially being more proactive in accessing health services as seen in prior research done on gender dynamics on health seeking behaviour (Thompson et al., 2016). Culturally and socially, women often bear the responsibility for family health and may be more engaged in community health programs such as those in

antimicrobial stewardship (Assai et al., 2006). Additionally, prior research shows that women have a higher incidence of E. coli infections (53%) compared to males which could increase their representation in health-related studies(Bennett et al., 1995).

Despite the predominance of female clients (67%, 1040) compared to (33%, 502) males in the sample, a slightly higher proportion of antimicrobial resistance was observed among male patients (56.2%) than female patients (52.4%). This finding suggests that while women were more represented in the study, men may be at a relatively higher risk of harbouring resistant strains. This is in line with previous research studies which report higher proportions of resistant E. coli in males compared to females (Khanal et al., 2024; McGregor et al., 2013). There are differences in antibiotic use patterns, occupational exposures, or underlying health conditions among men that could contribute to the observed resistance disparity (Batheja et al., 2025; Brandl et al., 2021). This sex-based variation in resistance highlights the importance of tailoring AMR interventions to address gender-specific behaviours and risks (Batheja et al., 2025). However, since the magnitude of these differences was generally less than 5%, this means the results may not represent clinically meaningful differences (McGregor et al., 2013).

The average and median ages were notably higher among male clients (mean = 42.8 years; median = 42 years) compared to female clients (mean = 32.9 years; median = 30 years), suggesting that older men were more likely to seek care or be captured in the sample. This age disparity may reflect gendered patterns in health-seeking behaviour, with men possibly delaying care until later in life or presenting with more complex conditions (Keene & Li, 2005; Nyalela et al., 2018). These demographic characteristics are important for interpreting antimicrobial resistance patterns, as age and sex may influence infection risk, antibiotic exposure, and treatment outcomes (Evans & Wiley, 2023).

7.2 Risk Factors for Drug Resistance

7.2.1 Feature importance and selection

Across the random forest, xgboost, and gradient boosting models, several features consistently emerged as strong predictors of antimicrobial resistance. The weighted importance score and frequency count framework aggregated feature importance scores from random forest, gbm, xgboost, and logistic regression to provide the top 7 predictors of drug resistance. Consistent with findings by Al Tawil et al, selecting between 5 and 15 well-justified features can optimize model performance while minimizing overfitting and maintaining interpretability, particularly in health-related predictive tasks (Al Tawil et al., 2024).

The number of days a patient had been on antibiotics emerged as the most influential variable in predicting resistance, highlighting the role of prior antibiotic exposure in shaping resistance outcomes. This finding aligns with existing evidence that prolonged or repeated antibiotic use promotes the emergence of drug resistant pathogens (Griskaitis et al., 2022). Clinically, this highlights the importance of monitoring treatment duration and avoiding unnecessary antibiotic courses, especially in outpatient settings in low-resource settings where empirical treatment is common (Okello et al., 2020; White et al., 2019).

The specific antibiotic tested on the E. coli-positive sample, was the second most influential variable, suggesting that certain antibiotics are more frequently associated with resistance patterns. This is consistent with research findings that demonstrate that antibiotic resistance patterns in E. coli vary significantly based on the specific antibiotics tested (Pouwels et al., 2019). Research found that amoxicillin prescribing was positively associated with resistance to multiple antibiotics including amoxicillin and ciprofloxacin among E. coli from urinary samples, indicating complex cross-resistance patterns (Pouwels et al., 2019). Similarly, another study identified tetracycline, cephalothin, sulfisoxazole, and streptomycin as antibiotics with the highest resistance frequencies among patients with E. Coli infections (Sayah et al., 2005). Clinically, this highlights the need for careful antibiotic selection based on local resistance profiles to avoid ineffective treatment and reduce drug resistance risks (López Romo & Quirós, 2019). From a public health perspective, it reinforces the importance of surveillance systems that track resistance trends by antibiotic type, enabling targeted stewardship interventions and informing national treatment guidelines (Diallo et al., 2020) .

Age ranked third, indicating that patient age is a key predictor for drug resistance. This is in line with prior research by Garcia A et al, Adam et al, and Yoshikawa which found that older patients were more likely to have resistant infections potentially due to cumulative antibiotic exposure or comorbidities (Adam et al., 2013; Garcia et al., 2017; Yoshikawa, 2002). From a public health perspective, it highlights the importance of tailoring AMR interventions to demographic risk profiles (Waterlow et al., 2025). Notably, a recent meta-analysis revealed that while antibiotic resistance probability generally increases with host age on average, diverse patterns exist including negative, humped, and U-shaped relationships depending on the specific bacteria-antibiotic combination (Binsted & McNally, 2024). Future research should explore these nuanced patterns to inform more targeted surveillance and treatment strategies.

The health facility where the sample was collected and processed was ranked forth suggesting that location-specific factors may influence resistance patterns. This variability agrees with prior research done in Uganda which demonstrates that Escherichia coli resistance patterns are influenced by location-

specific factors, including ecological, socioeconomic, and management practices (Weiss et al., 2018). This variation may partly be explained by the uneven distribution of resistance outcomes in the data collected across Regional Referral Hospitals (Mayito et al., 2024). Facilities with fewer resistant isolates may have yielded fewer stable estimates, potentially exaggerating or underrepresenting true prevalence. RRHs with a higher number of resistant outcomes likely provided more robust insights into local resistance patterns. However, while sample imbalance is a valid concern, this study assumes that it is also possible that the observed differences genuinely reflect true geographic variation in resistance. In this case, attempts to artificially balance the data could introduce bias and obscure meaningful location-specific insights. Therefore, while acknowledging the limitations posed by sample size variability, it is important to interpret these findings within the broader context of local prescribing practices, population dynamics, and healthcare infrastructure, which may all contribute to the observed resistance patterns.

Specimen type ranked fifth among predictors, indicating its relevance in shaping AMR patterns. Literature shows that specimen types such as blood, sputum, pus, and urine exhibit significant differences in resistance profiles (Wen et al., 2023). Notably, non-blood specimens often harbour higher resistance rates. For example, tracheal aspirates and pus swabs common in surgical and ICU setting have shown resistance rates ranging from 40–100% across multiple antibiotic classes (Mboowa et al., 2021; Moolchandani, 2017). From a public health perspective, expanding surveillance to include high-resistance non-priority specimens may improve detection of drug resistant organisms in critical care settings.

A patient's history of being hospitalized for more than 48 hours was the sixth most influential predictor of resistance, likely reflecting increased exposure to hospital-acquired pathogens and invasive procedures (Fridkin, 2001; Sommerstein et al., 2018) . From a public health perspective, it underscores the importance of monitoring resistance trends in healthcare-associated infections (Fridkin, 2001).

Prior exposure to antibiotics before sample collection ranked seventh as a predictor of resistance. This supports existing evidence that previous antibiotic use impact resistance outcome, especially when treatment is empirical or incomplete (Dryden et al., 2011). Public health strategies should therefore prioritize education on appropriate antibiotic use and strengthen prescription oversight to reduce unnecessary exposure.

Interestingly, duration on antibiotics which ranked first in xgboost and random forests, was the least important feature in logistic regression, while the specific antibiotic being tested on a sample dropped to eighth place. This discrepancy may be attributed to the linear nature of logistic regression, which may not capture complex interactions between variables as effectively as tree-based models (Halloran, 2009). A similar study found that logistic regression consistently underperformed across all metrics

compared to tree-based models like xgboost and Lightgbm, which demonstrates that logistic regression often underperforms compared to tree-based and ensemble methods when complex variable interactions are present (Pratiwi et al., 2024).

A novel finding in this study is the strong predictive power of the health facility where the sample was collected and processed in determining antimicrobial resistance outcomes. While geographic variation in resistance has been acknowledged in broader regional or national studies, few have incorporated facility-level data into predictive models for this study setting. The consistent ranking of the health facility where the sample was collected and processed among the top features in ensemble models suggests that hospital-specific factors such as prescribing practices, diagnostic protocols, and local microbial ecology may significantly influence resistance patterns (Li et al., 2025). This insight highlights the importance of integrating granular, site-level data into AMR surveillance frameworks, especially in settings with diverse healthcare infrastructures like Uganda.

7.2.2 Exploring Variable Relationships Using a Correlation Matrix

The correlation analysis revealed that most variables exhibited weak correlations (values close to 0), indicating that they likely provide independent information for modelling purposes. This independence is beneficial for machine learning, as it reduces redundancy and enhances the model's ability to learn distinct patterns from each feature. Notably, the duration of antibiotic use showed a positive correlation with both a patient's history of prior antibiotic intake and hospitalisation for more than 48 hours, suggesting that patients with such histories tend to be on antibiotics for longer periods. Conversely, there was a moderate negative correlation between duration of antibiotic use and specimen type, implying that certain specimen types may be associated with shorter or longer treatment durations.

Additionally, a moderate positive correlation was observed between hospitalisation history and prior antibiotic therapy, which aligns with clinical expectations, patients who have been hospitalised are more likely to have received antibiotics previously. Importantly, the absence of very strong correlations ($r > 0.8$) among variables reduces concerns about multicollinearity, a condition that can distort model estimates and reduce interpretability. This suggests that most variables can be retained in machine learning models without the need for dimensionality reduction or feature elimination due to redundancy. These insights justify the use of ensemble models that can capture complex, non-linear interactions among variables and highlight the importance of integrating these predictors into clinical decision-support tools (Al Musyaffa et al., 2025).

7.2.3 Triangulation of the top 7 features using the Weighted Importance Score and Frequency Count framework

The resulting WISFC scores provide a consensus-based ranking that supports more reliable interpretation. As visualized in the bar chart, the seven top-ranked features selected for model development were: the antibiotic to be tested on a sample or given to the patient, the age of the patient, the health facility where the sample was collected and processed, the specimen type, the duration the patient has spent on antibiotics (in days), patient history of hospitalization, and patient history of prior antibiotic therapy. These findings highlight that both clinical and contextual factors significantly influence antibiotic resistance patterns. For instance, patient age and hospitalization history may reflect underlying comorbidities or exposure to resistant strains, while facility-level differences could indicate variations in prescribing practices or infection control measures. The prominence of prior antibiotic use underscores its role as a key driver of resistance, reinforcing the need for stewardship interventions. Collectively, these features provide a strong foundation for predictive modeling, enabling more targeted and informed decision-making in antimicrobial therapy.

7.2 Model Performance

XGBoost achieved the highest performance across all evaluation metrics demonstrating strong predictive power and generalizability. Its ROC AUC of 90% indicates excellent discrimination between resistant and susceptible cases. In prior research, this superior performance can be attributed to XGBoost's ability to handle complex, non-linear relationships as well as its use of regularization toprevent overfitting (Shrivastava et al., 2024). These models also handle missing values and categorical variables more efficiently, making them well-suited for real-world clinical datasets that often contain noise and inconsistencies. In contrast, the decision tree model showed the weakest performance, highlighting its limitations in capturing complex, non-linear relationships. This is consistent with research which has demonstrated ensemble machine learning methods as more effective than decision trees for predicting antibiotic resistance in E. coli (Al Musyaffa et al., 2025).

The confusion matrices reveal differences in predictive performance across the four models. Lightgbm and xgboost outperform the others, showing the highest number of correctly predicted resistant cases and the lowest wrongly classified resistant cases, which is important in antimicrobial resistance prediction where missing resistant cases can lead to inappropriate treatment. Decision Tree, while simpler, shows the weakest performance with higher misclassification rates. Gradient boosting improves upon decision tree but still lags behind xgboost and Lightgbm. These results suggest that ensemble methods, particularly xgboost and lightgbm, are better suited for AMR prediction tasks due to their ability to capture complex patterns and reduce error rates. Their superior performance in identifying resistant cases supports their integration into clinical decision-support systems, where

accurate classification can guide more effective antibiotic prescribing and reduce the risk of treatment failure.

It is important to note that the model outputs a probability score rather than a definitive classification, leaving room for uncertainty in predictions. L. Wynants et al. reported that probabilistic approach allows clinicians to set thresholds for decision-making, but it also means that borderline cases may be misclassified depending on the chosen cutoff (Wynants et al., 2019). Such uncertainty can influence treatment choices and resource allocation, underscoring the need for careful calibration of thresholds and integration with clinical judgment to minimize risks (Wynants et al., 2019).

7.3 Development of the Interface

Streamlit and Python have consistently been adopted in scientific and medical research for developing user-friendly, interactive web applications (Keerthi et al., 2023; Monks & Harper, 2023; Müller et al., 2025). Python offers a rich ecosystem of libraries for data analysis and machine learning, while Streamlit simplifies the deployment of predictive models into intuitive interfaces without requiring extensive web development skills (Bergstra et al., 2015; Monks & Harper, 2023). Together, they enable rapid prototyping and real-time interaction, allowing users to input health data and instantly receive feedback, often accompanied by graphical visualizations and dynamic outputs (Keerthi et al., 2023). This combination is particularly advantageous in clinical settings where usability, speed, and transparency are critical (Douze et al., 2025).

A key consideration for the practical implementation of this tool is its placement within the diagnostic process. The predictive interface is designed to be utilized after pathogen identification. This timing ensures that clinicians receive preliminary insights during the waiting period for AST results, which often takes 24 - 72 hours in resource-limited settings. By providing early predictions, the tool can guide empirical therapy more effectively, reduce inappropriate antibiotic use, and prioritize antibiotics for AST, thereby optimizing laboratory resources. This integration complements existing workflows without replacing AST, reinforcing its role as a decision-support tool rather than a diagnostic substitute.

Unlike generic implementations, this application is tailored for antimicrobial resistance (AMR) prediction in resource-limited settings. It integrates a high-performing XGBoost model trained on Ugandan surveillance data, incorporating site-specific, specimen-type, and patient-level factors to deliver context-specific predictions that accurately reflect local resistance patterns. The interface goes beyond simple prediction by offering actionable insights: medical doctors can view alternative

antibiotics for empirical prescribing, while laboratory staff can prioritize those antibiotics for susceptibility testing. Additionally, the tool includes interactive features such as confidence scores and feature importance explanations, promoting transparency and trust in machine learning outputs.

7.4 Implications for Clinical Practice and Policy

The integration of machine learning, specifically the XGBoost driven predictive model developed in this study, into clinical workflows presents a transformative opportunity for managing Escherichia coli infections, particularly in resource-limited settings like Uganda. By providing decision support during prescribing processes during waiting periods and in the absence of ASTs, this approach reduces reliance on empirical antibiotic therapy. This predictive tool offers clinicians actionable insights into likely resistance patterns based on patient demographics, clinical history, and specimen type, allowing for more informed antibiotic selection even before AST results are available.

In the Ugandan context, where laboratory reagents and resources are scarce, the model can help prioritize which antibiotics to test for susceptibility. By identifying the most promising treatment options, it supports rational use of resources and ensures that testing focuses on antibiotics with the highest likelihood of effectiveness. Ultimately, this approach has the potential to reduce the healthcare economic burden associated with antimicrobial resistance by improving treatment accuracy and resource utilization. By integrating such a predictive model into Uganda's e-health information management system (EHIMS), policymakers can enable targeted antibiotic testing, minimize waste of laboratory reagents, and reduce costs linked to prolonged hospital stays and ineffective treatments.

7.5 Limitations

Despite the strong performance of XGBoost, several limitations must be acknowledged. One key limitation of this study was the missingness in certain variables, notably duration on antibiotics, which turned out to be one of the most predictive features. To address this, missing values were imputed using predictive modelling techniques; however, this may still introduce bias or uncertainty in the final model. Additionally, even though utility of all variables was determined during the conceptualization phase based on their relevance to the research objectives, during data processing, some variables exhibited excessive missingness, posing a risk of bias and reducing the robustness of the analysis. For this reason, despite their conceptual importance, these variables, such as the diagnosis variable, were be excluded after careful consideration to maintain the validity and integrity of the findings. This exclusion limits the model's ability to account for clinical context and potentially reducing its predictive accuracy.

Although the model demonstrates high accuracy, misclassifying resistant cases as susceptible can lead to inappropriate antibiotic prescriptions, resulting in treatment failure, prolonged hospital stays, and increased financial burden on both patients and healthcare systems. Predicting susceptibility as resistance may prompt unnecessary use of expensive broad-spectrum antibiotics, accelerating antimicrobial resistance. These risks highlight that even small error rates can have significant clinical and economic consequences, emphasizing the inherent limitations of relying solely on machine learning models even when thresholds are carefully calibrated and models are integrated with clinician oversight and other diagnostic tools prior to real-world deployment.Another limitation in this study stems from disparities in the dataset, for example the disparities in the data collected at the different site which may affect the model's ability to generalize across diverse population groups, potentially skewing predictions or masking important regional trends. This highlights the need for more balanced and representative datasets in future research to ensure equitable and accurate predictive modelling.

CHAPTER 8: CONCLUSION AND RECOMMENDATIONS

8.1 Conclusion

Ensemble methods, particularly xgboost outperformed traditional models, achieving high accuracy, precision, F1 score recall, and AUC scores, thereby validating their suitability for AMR prediction tasks. The integration of the XGBoost model into a web-based interface demonstrates the potential for practical utility in clinical settings, enabling clinicians to make timely, data-driven decisions even in the absence of immediate antimicrobial susceptibility testing (AST) results. This tool supports targeted antibiotic selection, reduces reliance on empirical broad-spectrum therapy, and promotes rational use of limited laboratory resources.

Importantly, the study highlights the previously underexplored predictors such as specimen type and facility-level factors, emphasizing the need for context-specific modelling approaches in AMR surveillance. The inclusion of non-GLASS specimens in this study provides a wholesome outlook on the influence of the specimen type being tested on the possibility of the outcome being either resistant or non-resistant.

Furthermore, this study's approach of applying machine learning in this context aligns with SDG 9 (Industry, Innovation, and Infrastructure) by fostering innovation in healthcare delivery and encouraging the integration of data-driven approaches into public health systems. For low-resource settings, where traditional surveillance methods may be costly or logistically challenging, predictive analytics can enhance early detection, guide stewardship efforts, and inform policy, contributing to more resilient and responsive health systems.

8.2 Recommendations

This study serves as a proof of concept that demonstrates the potential of machine learning (ML) to predict drug resistance in Escherichia coli infections using demographic, clinical, and microbiological data collected from ten tertiary healthcare facilities in Uganda. While the model demonstrates strong performance on the internal dataset, external validation is essential to confirm its applicability in different clinical settings and populations. This is derived from the experience from the study by S. Bleeker et al where a diagnostic prediction model showed an impressive 0.83 area under the curve internally, but collapsed to 0.57 when externally validated (Bleeker et al., 2003). This project could be tested on independent datasets from other hospitals, regions, or time periods to ensure it captures diverse

resistance patterns and data variability. Without external validation, there is a risk of overestimating the model's real-world performance due to dataset-specific biases. Successful external validation would strengthens confidence in the model's reliability and supports its integration into clinical workflows. With the rollout of the e-health information management system (eHIMS) in public hospitals, deploying the Streamlit-based interface can leverage this infrastructure, ensuring smooth integration with digital workflows and minimizing reliance on standalone systems. Insights from these testing sites can inform iterative improvements and support broader adoption across the health system.

Given that the model already incorporates key predictors, the Ministry of Health (MoH) could pilot and refine this tool in alignment with national treatment guidelines, with the goal of embedding it into national AMR stewardship programs. Clinicians could be trained to interpret model outputs and use these insights to inform antibiotic selection during prescribing thereby improving treatment accuracy and reducing inappropriate antibiotic use. Public health officials can leverage aggregated model outputs to monitor resistance trends enabling data-driven resource allocation and targeted interventions.

The Ministry of Health should prioritize strengthening digital infrastructure to support the deployment and sustainability of web-based clinical decision-support tools. This includes investing in reliable internet connectivity, secure data storage systems, and user-friendly platforms that can be integrated into existing healthcare workflows. Enhanced infrastructure will improve accessibility for healthcare providers across different regions and ensure the long-term usability of these tools, ultimately supporting evidence-based decision-making and improving patient outcomes.

Future researchers are encouraged to build on this work by exploring additional pathogens to broaden its applicability. To strengthen future iterations of similar predictive models, researchers could explore strategies for integrating more comprehensive data and additional clinically relevant variables such as sex, admission status, and referral status. While this study focused on variables with the highest predictive power and data completeness to ensure model reliability, incorporating these factors in future work may enhance interpretability and broaden applicability across diverse patient contexts. They should also validate the model in diverse populations to ensure its generalizability.Researchers should assess the long-term impact of AI-assisted prescribing on antimicrobial resistance rates and patient outcomes to determine its effectiveness and sustainability. Embedding such tools into routine care represents a promising step toward more data-driven, context-specific antimicrobial stewardship in Uganda and similar settings.

REFERENCES

Adam, H. J., Baxter, M. R., Davidson, R. J., Rubinstein, E., Fanella, S., Karlowsky, J. A., Lagace-Wiens, P. R. S., Hoban, D. J., Zhanel, G. G., Zhanel, G. G., Hoban, D. J., Adam, H. J., Karlowsky, J. A., Baxter, M. R., Nichol, K. A., Lagace-Wiens, P. R. S., & Walkty, A. (2013). Comparison of pathogens and their antimicrobial resistance patterns in paediatric, adult and elderly patients in Canadian hospitals. *Journal of Antimicrobial Chemotherapy*, *68*(suppl 1), i31–i37. https://doi.org/10.1093/jac/dkt024

Al Musyaffa, A. R., Pristyanto, Y., & Mauliza, N. (2025). COMPARISON OF ENSEMBLE METHODS FOR DECISION TREE MODELS IN CLASSIFYING E. COLI BACTERIA. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(3), 514–522. https://doi.org/10.33480/jitk.v10i3.5972

Al Tawil, A., Almazaydeh, L., Alqudah, B., Zaid Abualkishik, A., & A. Alwan, A. (2024). Predictive modeling for breast cancer based on machine learning algorithms and features selection methods. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(2), 1937. https://doi.org/10.11591/ijece.v14i2.pp1937-1947

Andretta, M., Tavares, R. de M., Fusieger, A., Yamatogi, R. S., & Nero, L. A. (2024). Agreement of methods to assess antimicrobial susceptibility using *Escherichia coli* isolates as target models. *Letters in Applied Microbiology*, *77*(2). https://doi.org/10.1093/lambio/ovae009

Assai, M., Siddiqi, S., & Watts, S. (2006). Tackling social determinants of health through community based initiatives. *BMJ*, *333*(7573), 854–856. https://doi.org/10.1136/bmj.38988.607836.68

Babirye, S. R., Nsubuga, M., Mboowa, G., Batte, C., Galiwango, R., & Kateete, D. P. (2024). Machine learning-based prediction of antibiotic resistance in Mycobacterium tuberculosis clinical isolates from Uganda. *BMC Infectious Diseases*, *24*(1), 1391. https://doi.org/10.1186/s12879-024-10282-7

Batheja, D., Goel, S., & Charani, E. (2025). Understanding gender inequities in antimicrobial resistance: role of biology, behaviour and gender norms. *BMJ Global Health*, *10*(1), e016711. https://doi.org/10.1136/bmjgh-2024-016711

Bello, R. H., Ibrahim, Y. K. E., Olayinka, B. O., Jimoh, A. A. G., Afolabi-Balogun, N. B., Oni-Babatunde, A. O., Olabode, H. O. K., David, M. S., Aliyu, A., & Olufadi - Ahmed, H. Y. (2021). Molecular Characterization of Extended Spectrum Beta – Lactamase Producing Escherichia Coli Isolated from Pregnant Women with Urinary Tract Infections Attending Ante–Natal Clinics in Ilorin Metropolis. *Nigerian Journal of Pharmaceutical Research*, *17*(1), 119–129. https://doi.org/10.4314/njpr.v17i1.13

Bender, J. B., Smith, K. E., McNees, A. A., Rabatsky-Ehr, T. R., Segler, S. D., Hawkins, M. A., Spina, N. L., Keene, W. E., Kennedy, M. H., Van Gilder, T. J., & Hedberg, C. W. (2004). Factors

Affecting Surveillance Data on *Escherichia coli* O157 Infections Collected from FoodNet Sites, 1996–1999. *Clinical Infectious Diseases*, *38*(s3), S157–S164. https://doi.org/10.1086/381582

Bennett, C. J., Young, M. N., & Darrington, H. (1995). Differences in urinary tract infections in male and female spinal cord injury patients on intermittent catheterization. *Spinal Cord*, *33*(2), 69–72. https://doi.org/10.1038/sc.1995.17

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.*, *13*, 281–305. https://api.semanticscholar.org/CorpusID:15700257

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 014008. https://doi.org/10.1088/1749-4699/8/1/014008

Binsted, L. E., & McNally, L. (2024). *Diverse Relationships Between Antibiotic Resistance and Host Age: A Meta-Analysis Across Antibiotic Classes and Bacterial Genera*. https://doi.org/10.1101/2024.02.25.24303263

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. M. (2003). External validation is necessary in prediction research: *Journal of Clinical Epidemiology*, *56*(9), 826–832. https://doi.org/10.1016/S0895-4356(03)00207-5

Brandl, M., Hoffmann, A., Willrich, N., Reuss, A., Reichert, F., Walter, J., Eckmanns, T., & Haller, S. (2021). Bugs That Can Resist Antibiotics but Not Men: Gender-Specific Differences in Notified Infections and Colonisations in Germany, 2010–2019. *Microorganisms*, *9*(5), 894. https://doi.org/10.3390/microorganisms9050894

Canbek, G., Temizel, T. T., & Sagiroglu, S. (2021). TasKar: A Research and Education Tool for Calculation and Representation of Binary Classification Performance Instruments. *2021 International Conference on Information Security and Cryptology (ISCTURKEY)*, 105–110. https://doi.org/10.1109/ISCTURKEY53027.2021.9654359

Chung, H. C., Foxx, C. L., Hicks, J. A., Stuber, T. P., Friedberg, I., Dorman, K. S., & Harris, B. (2023). An accurate and interpretable model for antimicrobial resistance in pathogenic Escherichia coli from livestock and companion animal species. *PloS One*, *18*(8), e0290473. https://doi.org/10.1371/journal.pone.0290473

Daneman, N., Fridman, D., Johnstone, J., Langford, B. J., Lee, S. M., MacFadden, D. M., Mponponsuo, K., Patel, S. N., Schwartz, K. L., & Brown, K. A. (2023). Antimicrobial resistance and mortality following E. coli bacteremia. *EClinicalMedicine*, *56*, 101781. https://doi.org/10.1016/j.eclinm.2022.101781

Denny, K. J., Cotta, M. O., Parker, S. L., Roberts, J. A., & Lipman, J. (2016). The use and risks of antibiotics in critically ill patients. *Expert Opinion on Drug Safety*, *15*(5), 667–678. https://doi.org/10.1517/14740338.2016.1164690

Diallo, O. O., Baron, S. A., Abat, C., Colson, P., Chaudet, H., & Rolain, J.-M. (2020). Antibiotic resistance surveillance systems: A review. *Journal of Global Antimicrobial Resistance*, *23*, 430–438. https://doi.org/10.1016/j.jgar.2020.10.009

Douze, L., Schiro, J., & Pelayo, S. (2025). *Integrate Usability Evaluation into Clinical Investigation: Why, When and How?* https://doi.org/10.3233/SHTI250236

Dryden, M., Johnson, A. P., Ashiru-Oredope, D., & Sharland, M. (2011). Using antibiotics responsibly: right drug, right time, right dose, right duration. *Journal of Antimicrobial Chemotherapy*, *66*(11), 2441–2443. https://doi.org/10.1093/jac/dkr370

European Committee on Antimicrobial Susceptibility Testing (EUCAST). (2025). *EUCAST Clinical Breakpoints Version 2.0–5.0*. Https://Www.Eucast.Org/Ast_of_bacteria. https://www.eucast.org/ast_of_bacteria

Feng, L., Wu, H., Yue, H., Chu, Y., Zhang, J., Huang, X., Pang, S., Zhang, L., Li, Y., Wang, W., Zou, B., & Zhou, G. (2022). Multiplexed and Rapid AST for *Escherichia coli* Infection by Simultaneously Pyrosequencing Multiple Barcodes Each Specific to an Antibiotic Exposed to a Sample. *Analytical Chemistry*, *94*(24), 8633–8641. https://doi.org/10.1021/acs.analchem.2c00312

Fridkin, S. K. (2001). Increasing prevalence of antimicrobial resistance in intensive care units. *Critical Care Medicine*, *29*(Supplement), N64–N68. https://doi.org/10.1097/00003246-200104001-00002

Garcia, A., Delorme, T., & Nasr, P. (2017). Patient age as a factor of antibiotic resistance in methicillin-resistant Staphylococcus aureus. *Journal of Medical Microbiology*, *66*(12), 1782–1789. https://doi.org/10.1099/jmm.0.000635

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. https://api.semanticscholar.org/CorpusID:7467506

Griskaitis, M., Furuya-Kanamori, L., Allel, K., Stabler, R., Harris, P., Paterson, D. L., & Yakob, L. (2022). β-Lactam–Resistant *Streptococcus pneumoniae* Dynamics Following Treatment: A Dose-Response Meta-analysis. *Clinical Infectious Diseases*, *75*(11), 1962–1970. https://doi.org/10.1093/cid/ciac293

Gurung, D., Bhardwaj, V. K., & Fotedar, S. (2024). Antimicrobial Resistance Challenge to Sustainable Development Goals and Targets: A One-Health Perspective. *Journal of Integrative Medicine and Public Health*, *3*(1), 9–13. https://doi.org/10.4103/JIMPH.JIMPH_11_24

Halloran, J. T. (2009). *Classification: Naive Bayes vs Logistic Regression*. https://api.semanticscholar.org/CorpusID:63078784

Hope, M., Kiggundu, R., Tabajjwa, D., Tumwine, C., Lwigale, F., Mwanja, H., Waswa, J. P., Mayito, J., Bulwadda, D., Byonanebye, D. M., Kakooza, F., & Kambugu, A. (2024). Progress on implementing the WHO-GLASS recommendations on priority pathogen-antibiotic sensitivity testing in Africa: A scoping review. *Wellcome Open Research*, *9*, 692. https://doi.org/10.12688/wellcomeopenres.23133.1

Hur, R., Golik, S., & She, Y. (2024). Leveraging Large Data, Statistics, and Machine Learning to Predict the Emergence of Resistant E. coli Infections. *Pharmacy*, *12*(2), 53. https://doi.org/10.3390/pharmacy12020053

Hu, Y., Matsui, Y., & W. Riley, L. (2020). Risk factors for fecal carriage of drug-resistant Escherichia coli: a systematic review and meta-analysis. *Antimicrobial Resistance & Infection Control*, *9*(1), 31. https://doi.org/10.1186/s13756-020-0691-3

Ingle, D. J., Levine, M. M., Kotloff, K. L., Holt, K. E., & Robins-Browne, R. M. (2018). Dynamics of antimicrobial resistance in intestinal Escherichia coli from children in community settings in South Asia and sub-Saharan Africa. *Nature Microbiology*, *3*(9), 1063–1073. https://doi.org/10.1038/s41564-018-0217-4

Kapisi, J., Sserwanga, A., Kitutu, F. E., Rutebemberwa, E., Awor, P., Weber, S., Keller, T., Kaawa-Mafigiri, D., Ekusai-Sebatta, D., Horgan, P., Dittrich, S., Moore, C. E., Salami, O., Olliaro, P., Nkeramahame, J., & Hopkins, H. (2023). Impact of the Introduction of a Package of Diagnostic Tools, Diagnostic Algorithm, and Training and Communication on Outpatient Acute Fever Case Management at 3 Diverse Sites in Uganda: Results of a Randomized Controlled Trial. *Clinical Infectious Diseases*, *77*(Supplement_2), S156–S170. https://doi.org/10.1093/cid/ciad341

Keene, J., & Li, X. (2005). Age and Gender Differences in Health Service Utilization. *Journal of Public Health*, *27*(1), 74–79. https://doi.org/10.1093/pubmed/fdh208

Keerthi, M. P., Reddy, G. S., Raghava, V. S., & Reddy, K. B. (2023). Streamlit Interface for Multiple Disease Diagnosis. *International Journal for Research in Applied Science and Engineering Technology*, *11*(2), 1159–1164. https://doi.org/10.22214/ijraset.2023.49166

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:3815895

Kherabi, Y., Thy, M., Bouzid, D., Antcliffe, D. B., Rawson, T. M., & Peiffer-Smadja, N. (2024). Machine learning to predict antimicrobial resistance: future applications in clinical practice? *Infectious Diseases Now*, *54*(3), 104864. https://doi.org/10.1016/j.idnow.2024.104864

Kiggundu, R., Lusaya, E., Seni, J., Waswa, J. P., Kakooza, F., Tjipura, D., Kikule, K., Muiva, C., Joshi, M. P., Stergachis, A., Kitutu, F. E., & Konduri, N. (2023). Identifying and addressing challenges to antimicrobial use surveillance in the human health sector in low- and middle-income countries: experiences and lessons learned from Tanzania and Uganda. *Antimicrobial Resistance & Infection Control*, *12*(1), 9. https://doi.org/10.1186/s13756-023-01213-3

Larramendy, S., Deglaire, V., Dusollier, P., Fournier, J.-P., Caillon, J., Beaudeau, F., & Moret, L. (2020). Risk Factors of Extended-Spectrum Beta-Lactamases-Producing Escherichia coli Community Acquired Urinary Tract Infections: A Systematic Review . *Infection and Drug Resistance*, *Volume 13*, 3945–3955. https://doi.org/10.2147/IDR.S269033

Li, H., Samore, M. H., & Zhang, Y. (2025). *Comparative Evaluation of Time Series Forecasting Approaches for Facility-Level Antibiotic Resistance Outcomes in the Veterans Health Administration*. https://doi.org/10.1101/2025.04.17.25325964

López Romo, A., & Quirós, R. (2019). Appropriate use of antibiotics: an unmet need. *Therapeutic Advances in Urology*, *11*. https://doi.org/10.1177/1756287219832174

MacFadden, D. R., Fisman, D., Andre, J., Ara, Y., Majumder, M. S., Bogoch, I. I., Daneman, N., Wang, A., Vavitsas, M., Castellani, L., & Brownstein, J. S. (2016). A Platform for Monitoring Regional Antimicrobial Resistance, Using Online Data Sources: ResistanceOpen. *The Journal of Infectious Diseases*, *214*(suppl_4), S393–S398. https://doi.org/10.1093/infdis/jiw343

MacIntyre, C. R., & Bui, C. M. (2017). Pandemics, public health emergencies and antimicrobial resistance - putting the threat in an epidemiologic and risk analysis context. In *Archives of Public Health* (Vol. 75, Issue 1). https://doi.org/10.1186/s13690-017-0223-7

MacKinnon, M. C., Sargeant, J. M., Pearl, D. L., Reid-Smith, R. J., Carson, C. A., Parmley, E. J., & McEwen, S. A. (2020). Evaluation of the health and healthcare system burden due to antimicrobial-resistant Escherichia coli infections in humans: a systematic review and meta-analysis. *Antimicrobial Resistance & Infection Control*, *9*(1), 200. https://doi.org/10.1186/s13756-020-00863-x

Martens, E., & Demain, A. L. (2017). The antibiotic resistance crisis, with a focus on the United States. *The Journal of Antibiotics*, *70*(5), 520–526. https://doi.org/10.1038/ja.2017.30

Mayito, J., Kibombo, D., Olaro, C., Nabadda, S., Guma, C., Nabukenya, I., Busuge, A., Dhikusooka, F., Andema, A., Mukobi, P., Onyachi, N., Watmon, B., Obbo, S., Yayi, A., Elima, J., Barigye, C., Nyeko, F. J., Mugerwa, I., Sekamatte, M., … Kajumbula, H. (2024). Characterization of Antibiotic Resistance in Select Tertiary Hospitals in Uganda: An Evaluation of 2020 to 2023 Routine Surveillance Data. *Tropical Medicine and Infectious Disease*, *9*(4), 77. https://doi.org/10.3390/tropicalmed9040077

Mboowa, G., Aruhomukama, D., Sserwadda, I., Kitutu, F. E., Davtyan, H., Owiti, P., Kamau, E. M., Enbiale, W., Reid, A., Bulafu, D., Kisukye, J., Lubwama, M., & Kajumbula, H. (2021). Increasing Antimicrobial Resistance in Surgical Wards at Mulago National Referral Hospital, Uganda, from 2014 to 2018—Cause for Concern? *Tropical Medicine and Infectious Disease*, *6*(2), 82. https://doi.org/10.3390/tropicalmed6020082

McEwen, S. A., & Collignon, P. J. (2018). Antimicrobial Resistance: a One Health Perspective. *Microbiology Spectrum*, *6*(2). https://doi.org/10.1128/microbiolspec.arba-0009-2017

McGregor, J. C., Elman, M. R., Bearden, D. T., & Smith, D. H. (2013). Sex- and age-specific trends in antibiotic resistance patterns of Escherichia coli urinary isolates from outpatients. *BMC Family Practice*, *14*(1), 25. https://doi.org/10.1186/1471-2296-14-25

Mitrani-Gold, F. S., Kaye, K. S., Gupta, V., Mulgirigama, A., Trautner, B. W., Scangarella-Oman, N. E., Yu, K. C., Ye, G., & Joshi, A. V. (2023). Older patient age and prior antimicrobial use strongly

predict antimicrobial resistance in Escherichia coli isolates recovered from urinary tract infections among female outpatients. *PLOS ONE*, *18*(5), e0285427. https://doi.org/10.1371/journal.pone.0285427

Monks, T., & Harper, A. (2023). Improving the usability of open health service delivery simulation models using Python and web apps. *NIHR Open Research*, *3*, 48. https://doi.org/10.3310/nihropenres.13467.2

Moolchandani, K. (2017). Antimicrobial Resistance Surveillance among Intensive Care Units of a Tertiary Care Hospital in South India. *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*. https://doi.org/10.7860/JCDR/2017/23717.9247

Moran, E., Robinson, E., Green, C., Keeling, M., & Collyer, B. (2020). Towards personalized guidelines: using machine-learning algorithms to guide antimicrobial selection. *Journal of Antimicrobial Chemotherapy*, *75*(9). https://doi.org/10.1093/jac/dkaa222

Müller, T. D., Siraj, A., Walter, A., Kim, J., Wein, S., von Kleist, J., Feroz, A., Pilz, M., Jeong, K., Sing, J. C., Charkow, J., Röst, H. L., & Sachsenberg, T. (2025). OpenMS WebApps: Building User-Friendly Solutions for MS Analysis. *Journal of Proteome Research*, *24*(2), 940–948. https://doi.org/10.1021/acs.jproteome.4c00872

Nabadda, S., Kakooza, F., Kiggundu, R., Walwema, R., Bazira, J., Mayito, J., Mugerwa, I., Sekamatte, M., Kambugu, A., Lamorde, M., Kajumbula, H., & Mwebasa, H. (2021). Implementation of the World Health Organization Global Antimicrobial Resistance Surveillance System in Uganda, 2015-2020: Mixed-Methods Study Using National Surveillance Data. *JMIR Public Health and Surveillance*, *7*(10), e29954. https://doi.org/10.2196/29954

Nadimpalli, M., Delarocque-Astagneau, E., Love, D. C., Price, L. B., Huynh, B.-T., Collard, J.-M., Lay, K. S., Borand, L., Ndir, A., Walsh, T. R., Guillemot, D., Borand, L., De Lauzanne, A., Kerleguer, A., Tarantola, A., Piola, P., Chon, T., Lach, S., Ngo, V., … Abdou, A. Y. (2018). Combating Global Antibiotic Resistance: Emerging One Health Concerns in Lower- and Middle-Income Countries. *Clinical Infectious Diseases*, *66*(6), 963–969. https://doi.org/10.1093/cid/cix879

Naghavi, M., Vollset, S. E., Ikuta, K. S., Swetschinski, L. R., Gray, A. P., Wool, E. E., Robles Aguilar, G., Mestrovic, T., Smith, G., Han, C., Hsu, R. L., Chalek, J., Araki, D. T., Chung, E., Raggi, C., Gershberg Hayoon, A., Davis Weaver, N., Lindstedt, P. A., Smith, A. E., … Murray, C. J. L. (2024). Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, *404*(10459), 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1

Nair, D., Navneethapandian, P. D., Tripathy, J. P., Harries, A. D., Klinton, J. S., Watson, B., Sivaramakrishnan, G. N., Reddy, D. S., Murali, L., Natrajan, M., & Swaminathan, S. (2016). Impact of rapid molecular diagnostic tests on time to treatment initiation and outcomes in patients with multidrug-resistant tuberculosis, Tamil Nadu, India. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, *110*(9), 534–541. https://doi.org/10.1093/trstmh/trw060

Nakate, M. G., Moleki, M., Sarki, A., & Fleming, V. (2022). Health Workers' Documentation Process as a Prerequisite to the Integration of Patient Care at a Regional Referral Hospital in Uganda. *Open Journal of Nursing*, *12*(09), 616–632. https://doi.org/10.4236/ojn.2022.129042

Nedungadi, P., Surendran, S., Tang, K.-Y., & Raman, R. (2024). Big Data and AI Algorithms for Sustainable Development Goals: A Topic Modeling Analysis. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2024.3516500

Nkansa-Gyamfi, N. A., Kazibwe, J., Traore, D. A. K., & Nji, E. (2019). Prevalence of multidrug-, extensive drug-, and pandrug-resistant commensal *Escherichia coli* isolated from healthy humans in community settings in low- and middle-income countries: a systematic review and meta-analysis. *Global Health Action*, *12*(sup1), 1815272. https://doi.org/10.1080/16549716.2020.1815272

Nsubuga, M., Galiwango, R., Jjingo, D., & Mboowa, G. (2024). Generalizability of machine learning in predicting antimicrobial resistance in E. coli: a multi-country case study in Africa. *BMC Genomics*, *25*(1), 287. https://doi.org/10.1186/s12864-024-10214-4

Nuwamanya, E., Mackline, H., & Kiggundu, R. (2025, April 4). *CAMO-Net presents economic burden of AMR to Uganda's Parliament*. https://camonet.org/2025/04/04/policy-in-action-camo-net-presents-economic-burden-of-amr-to-ugandas-parliament/

Nyalela, M., Dlungwane, T., Taylor, M., & Nkwanyana, N. (2018). Health seeking and sexual behaviour of men presenting with sexually transmitted infections in two primary health care clinics in Durban. *Southern African Journal of Infectious Diseases*, 1–6. https://doi.org/10.1080/23120053.2018.1520480

Obakiro, S. B., Kiyimba, K., Paasi, G., Napyo, A., Anthierens, S., Waako, P., Royen, P. Van, Iramiot, J. S., Goossens, H., & Kostyanev, T. (2021). Prevalence of antibiotic-resistant bacteria among patients in two tertiary hospitals in Eastern Uganda. *Journal of Global Antimicrobial Resistance*, *25*, 82–86. https://doi.org/10.1016/j.jgar.2021.02.021

Okello, N., Oloro, J., Kyakwera, C., Kumbakumba, E., & Obua, C. (2020). Antibiotic prescription practices among prescribers for children under five at public health centers III and IV in Mbarara district. *PLOS ONE*, *15*(12), e0243868. https://doi.org/10.1371/journal.pone.0243868

Panda, P. K. (2025). Wrong diagnosis — Wrong antimicrobials: Rise in antimicrobial resistance in developing countries. *IDCases*, *41*, e02313. https://doi.org/10.1016/j.idcr.2025.e02313

Pathak, A., Chandran, S. P., Mahadik, K., Macaden, R., & Lundborg, C. S. (2013). Frequency and factors associated with carriage of multi-drug resistant commensal Escherichia coliamong women attending antenatal clinics in Central India. *BMC Infectious Diseases*, *13*(1), 199. https://doi.org/10.1186/1471-2334-13-199

Pouwels, K. B., Muller-Pebody, B., Smieszek, T., Hopkins, S., & Robotham, J. V. (2019). Selection and co-selection of antibiotic resistances among Escherichia coli by antibiotic use in primary care:

An ecological analysis. *PLOS ONE*, *14*(6), e0218134. https://doi.org/10.1371/journal.pone.0218134

Pratiwi, N. K. C., Tayara, H., & Chong, K. T. (2024). An Ensemble Classifiers for Improved Prediction of Native–Non-Native Protein–Protein Interaction. *International Journal of Molecular Sciences*, *25*(11), 5957. https://doi.org/10.3390/ijms25115957

Raj, A. (2019). A Review on Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *7*(6), 792–796. https://doi.org/10.22214/ijraset.2019.6138

Richter, L., Plessis, E. Du, Duvenage, S., & Korsten, L. (2021). High prevalence of multidrug resistant *Escherichia coli* isolated from fresh vegetables sold by selected formal and informal traders in the most densely populated Province of South Africa. *Journal of Food Science*, *86*(1), 161–168. https://doi.org/10.1111/1750-3841.15534

Riyanto, S., Sitanggang, I. S., Djatna, T., & Atikah, T. D. (2023). Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. *International Journal of Advanced Computer Science and Applications*, *14*(6). https://doi.org/10.14569/IJACSA.2023.01406116

Rosini, R., Nicchi, S., Pizza, M., & Rappuoli, R. (2020). Vaccines Against Antimicrobial Resistance. In *Frontiers in Immunology* (Vol. 11). https://doi.org/10.3389/fimmu.2020.01048

Sakagianni, A., Koufopoulou, C., Feretzakis, G., Kalles, D., Verykios, V. S., Myrianthefs, P., & Fildisis, G. (2023). Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. In *Antibiotics* (Vol. 12, Issue 3). https://doi.org/10.3390/antibiotics12030452

Samy, A. A., Mansour, A. S., Khalaf, D. D., & Khairy, E. A. (2022). Development of multidrug-resistant Escherichia coli in some Egyptian veterinary farms. *Veterinary World*, 488–495. https://doi.org/10.14202/vetworld.2022.488-495

Segawa, I., Ssebambulidde, K., Kiiza, D., & Mukonzo, J. (2020). *Antimicrobial Sensitivity Testing Using the Kirby-Bauer Disk Diffusion Method; Limited Utility in Ugandan Hospitals*. https://doi.org/10.31730/osf.io/jh96e

Shrivastava, A., Kotiyal, A., Habelalmateen, M. I., Rana, A., Devi, V. S. A., Rao, B. D., & Bansal, S. (2024). Leveraging XGBoost for Predictive Analytics in Healthcare: Enhancing Disease Diagnosis. *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 1666–1672. https://doi.org/10.1109/IC3I61595.2024.10829136

Singh*, K., & Mukherjee, A. (2022). Reliable Algorithms for Machine Learning Models: Implementation Research in Data Science. *International Journal of Recent Technology and Engineering (IJRTE)*, *10*(6), 102–106. https://doi.org/10.35940/ijrte.F6871.0310622

Sommerstein, R., Atkinson, A., Lo Priore, E. F., Kronenberg, A., Marschall, J., Burnens, A., Cherkaoui, A., Dubuis, O., Egli, A., Gaia, V., Koch, D., Kronenberg, A., Leib, S. L., Luyet, S., Nordmann, P., Perreten, V., Piffaretti, J.-C., Prod'hom, G., Schrenzel, J., … Zbinden, R. (2018).

Characterizing non-linear effects of hospitalisation duration on antimicrobial resistance in respiratory isolates: an analysis of a prospective nationwide surveillance system. *Clinical Microbiology and Infection*, *24*(1), 45–52. https://doi.org/10.1016/j.cmi.2017.05.018

Stalteri Mastrangelo, R., Santesso, N., Bognanni, A., Darzi, A., Karam, S., Piggott, T., Baldeh, T., Schünemann, F., Ventresca, M., Morgano, G. P., Moja, L., Loeb, M., & Schunemann, H. (2021). Consideration of antimicrobial resistance and contextual factors in infectious disease guidelines: a systematic survey. *BMJ Open*, *11*(7), e046097. https://doi.org/10.1136/bmjopen-2020-046097

Talekar, B. (2020). A Detailed Review on Decision Tree and Random Forest. *Bioscience Biotechnology Research Communications*, *13*(14), 245–248. https://doi.org/10.21786/bbrc/13.14/57

Tang, K. W. K., Millar, B. C., & Moore, J. E. (2023). Antimicrobial Resistance (AMR). *British Journal of Biomedical Science*, *80*. https://doi.org/10.3389/bjbs.2023.11387

Thompson, A. E., Anisimowicz, Y., Miedema, B., Hogg, W., Wodchis, W. P., & Aubrey-Bassler, K. (2016). The influence of gender and other patient characteristics on health care-seeking behaviour: a QUALICOPC study. *BMC Family Practice*, *17*(1), 38. https://doi.org/10.1186/s12875-016-0440-0

Tseng, W.-P., Chen, Y.-C., Yang, B.-J., Chen, S.-Y., Lin, J.-J., Huang, Y.-H., Fu, C.-M., Chang, S.-C., & Chen, S.-Y. (2017). Predicting Multidrug-Resistant Gram-Negative Bacterial Colonization and Associated Infection on Hospital Admission. *Infection Control & Hospital Epidemiology*, *38*(10), 1216–1225. https://doi.org/10.1017/ice.2017.178

Tuem, K. B., Gebre, A. K., Atey, T. M., Bitew, H., Yimer, E. M., & Berhe, D. F. (2018). Drug Resistance Patterns of *Escherichia coli* in Ethiopia: A Meta-Analysis. *BioMed Research International*, *2018*, 1–13. https://doi.org/10.1155/2018/4536905

Uganda Bureau of Statistics. (2024, December). *NATIONAL POPULATION AND HOUSING CENSUS 2024 FINAL REPORT*. Https://Www.Ubos.Org/Wp-Content/Uploads/2024/12/National-Population-and-Housing-Census-2024-Final-Report-Volume-1-Main.Pdf. https://www.ubos.org/wp-content/uploads/2024/12/National-Population-and-Housing-Census-2024-Final-Report-Volume-1-Main.pdf

Valavarasu, S., Sangu, Y., & Mahapatra, T. (2025). Prediction of antibiotic resistance from antibiotic susceptibility testing results from surveillance data using machine learning. *Scientific Reports*, *15*(1), 30509. https://doi.org/10.1038/s41598-025-14078-w

Varghese, S. B., & K.J., E. (2024). Gradient Edge: Advancing Predictive Modelling with Enhanced Gradient Boosting: A Multi-Dataset Approach. *International Journal of Computer Applications*, *186*(49), 1–6. https://doi.org/10.5120/ijca2024924160

Veledar, E., Zhou, L., Veledar, O., Gardener, H., Gutierrez, C. M., Romano, J. G., & Rundek, T. (2025). Synthesizing Explainability Across Multiple ML Models for Structured Data. *Algorithms*, *18*(6), 368. https://doi.org/10.3390/a18060368

Vermeulen, A. F. (2020). Supervised Learning: Using Labeled Data for Insights. In *Industrial Machine Learning* (pp. 63–136). Apress. https://doi.org/10.1007/978-1-4842-5316-8_4

Walker, M. M., Roberts, J. A., Rogers, B. A., Harris, P. N. A., & Sime, F. B. (2022). Current and Emerging Treatment Options for Multidrug Resistant Escherichia coli Urosepsis: A Review. In *Antibiotics* (Vol. 11, Issue 12). https://doi.org/10.3390/antibiotics11121821

Walker, R. A., Stucka, T. P., Ezzine, M., Vasiliev, K., Sebudde, R. K., Juzon, J. D., Bagarukayo, D. R. N. S., Chugunov, D., & de Andrade Falcão, N. (2019). *Economic Development and Human Capital in Uganda - A Case for Investing More in Education*. https://api.semanticscholar.org/CorpusID:191667497

Waterlow, N. R., Chandler, C. I. R., Cooper, B., Moore, C. E., Robotham, J. V, Sartorius, B., Sharland, M., & Knight, G. M. (2025). *Combining demographic shifts with age-based resistance prevalence: a modelling estimate of future AMR burden in Europe and implications for targets*. https://doi.org/10.1101/2025.03.20.25324297

Weiss, D., Wallace, R. M., Rwego, I. B., Gillespie, T. R., Chapman, C. A., Singer, R. S., & Goldberg, T. L. (2018). Antibiotic-Resistant Escherichia coli and Class 1 Integrons in Humans, Domestic Animals, and Wild Primates in Rural Uganda. *Applied and Environmental Microbiology*, *84*(21). https://doi.org/10.1128/AEM.01632-18

Wen, H., Xie, S., Liu, Y., Liang, Y., Zhang, P., Wang, X., & Li, J. (2023). Retrospective Analysis of Sensitivity Characteristics of Enterobacteriaceae: A Study Based on Specimen Types, Sex, and Age Bracket of Patients. *Infection and Drug Resistance*, *Volume 16*, 1753–1765. https://doi.org/10.2147/IDR.S401341

White, A. T., Clark, C. M., Sellick, J. A., & Mergenhagen, K. A. (2019). Antibiotic stewardship targets in the outpatient setting. *American Journal of Infection Control*, *47*(8), 858–863. https://doi.org/10.1016/j.ajic.2019.01.027

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., & Van Calster, B. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, *17*(1), 192. https://doi.org/10.1186/s12916-019-1425-3

Yoshikawa, T. T. (2002). Antimicrobial Resistance and Aging: Beginning of the End of the Antibiotic Era? *Journal of the American Geriatrics Society*, *50*(s7), 226–229. https://doi.org/10.1046/j.1532-5415.50.7s.2.x

Appendix A: Code Book

Variable order	VariableName	Variabletype	Description ofvariable	Codingvalues andlabels
1	Lab	select_one	Facilityname	1. Arua 2. Kabale 3. Mbale 4. Jinja 5. Gulu 6. Lira 7. Soroti 8. Masaka 9. Mbarara
2	Identificationnumber	text	PatientID	
3	Specimennumber	text	Lab specimennumber	
4	Organism	text	Organism/pathogenisolated	
5	Sex	select_one	PatientSex	10. Female 11. Male
6	Age	text	Age in completeyears	
7	Department	text	HospitalWard	
8	Collectiondate	date	Date of samplecollection	
9	Day	text	Day of samplecollection	
10	Month	text	Monthofsamplecollection	
11	Year	text	Year of samplecollection	
12	Specimentype	text	Sampletype	

13	Comment	text	Doctor'scomment	
14	Hospitalized48hrs	yesno	Wasthepatienthospitalized formore than 48 hours?	12. Yes 2No
			Wasthepatienttransferredf rom another facility?	
15	Transferredfromanotherfacilit	yesno		13. Yes 2No
16	Priorantibiotictherapy	yesno	Wasthepatientonantibiotic spriortoadmission?	
17	Durationofantibioticsindays	text	Ifyes,forhowlongwasthep atient on antibiotics?	
18	Dateofadmission	text	Date ofadmission.	
19	Diagnosis	text	Diagnosis	
20	AMC_ND20	select_one	Amoxacillin/clavulanicac i d	1-R 2-I 3-S
21	AMK_ND30	select_one	Amikacin	1-R 2-I 3-S
22	AMP_ND10	select_one	Ampicillin	1-R 2-I 3-S

23	AMX_ND25	select_one	Amoxacillin	1-R 2-I 3-S
24	ATM_ND30	select_one	Aztreonam	1-R 2-I 3-S
				1-R

25	AZM_ND15	select_one	Azithromycin15mcg	2-I 3-S
26	AZM_ND30	select_one	Azithromycin30mcg	1-R 2-I 3-S
27	CAC_ND30	select_one	CAC_ND30	1-R 2-I 3-S
28	CFM_ND5	select_one	Cefixime	1-R 2-I 3-S

29	CFR_ND30	select_one	CFR_ND30	1-R 2-I 3-S
30	CHL_ND30	select_one	Chloramphenicol	1-R 2-I 3-S
31	CIP_ND5	select_one	Ciprofloxacin	1-R 2-I 3-S
32	CLI_ND2	select_one	Clindamycin	1-R 2-I 3-S
33	CLO_ND5	select_one	CLO_ND5	1-R 2-I 3-S
34	CLR_ND15	select_one	CLR_ND15	1-R 2-I 3-S

35	CRO_ND30	select_one	Ceftriaxone	1-R 2-I 3-S
36	CXM_ND30	select_one	Cefuroxime	1-R 2-I 3-S
37	CZX_ND30	select_one	CZX_ND30	1-R 2-I 3-S
38	DOX_ND30	select_one	Doxycline	1-R 2-I 3-S
39	ERY_ND15	select_one	Erythromycin	1-R 2-I 3-S
				1-R

40	ETP_ND10	select_one	Ertapenem	2-I 3-S

41	FEP_ND30	select_one	Cefepime	1-R 2-I 3-S
42	FOX_ND30	select_one	Cefoxitin	1-R 2-I 3-S
43	GAT_ND5	select_one	GAT_ND5	1-R 2-I 3-S
44	GEN_ND10	select_one	Gentamycin 10mcg	1-R 2-I 3-S
45	IPM_ND10	select_one	Imipenem	1-R 2-I 3-S
46	MEM_ND10	select_one	Meropenem	1-R 2-I 3-S
				1-R 2-I

47	MET_ND5	select_one	Metronidazole	3-S

48	MFX_ND5	select_one	Moxifloxacin	1-R 2-I 3-S
49	NAL_ND30	select_one	Nalidixicacid	1-R 2-I 3-S
50	NIT_ND300	select_one	Nitrofurantoin	1-R 2-I 3-S
51	OFX_ND5	select_one	OFX_ND5	1-R 2-I 3-S
52	OXA_ND1	select_one	Oxacillin	1-R 2-I 3-S

53	PEF_ND10	select_one	PEF_ND10	1-R 2-I 3-S
54	PEN_ND10	select_one	Penicilling	1-R 2-I 3-S
55	PIP_ND100	select_one	Piperacillin	1-R 2-I 3-S
56	POP_ND300	select_one	POP_ND300	1-R 2-I 3-S
57	PRL_ND2	select_one	PRL_ND2	1-R 2-I 3-S
58	SMX_ND200	select_one	SMX_ND200	1-R 2-I 3-S

59	SXT_ND12	select_one	Sulfamethoxazoletrimethoprim	1-R 2-I 3-S
				1-R 2-I

60	TCY_ND30	select_one	Tetracycline	3-S
61	TEC_ND30	select_one	TEC_ND30	1-R 2-I 3-S
62	TMP_ND5	select_one	TMP_ND5	1-R 2-I 3-S
63	TZP_ND100	select_one	Piperacillin/tazobactam	1-R 2-I 3-S
64	VAN_ND30	select_one	Vancomycin	1-R 2-I 3-S

65	BM	select_one	BM	1-R 2-I 3-S
66	CTX	select_one	Cefotaxime	1-R 2-I 3-S
67	SAM	select_one	SAM	1-R 2-I 3-S
68	CTC	select_one	Cefotaxime/clavunate	1-R 2-I 3-S
69	CAZ	select_one	Ceftazidime	1-R 2-I 3-S
70	DOR	select_one	Doripenem	1-R 2-I 3-S

71	GEH	select_one	Gentamycin 120mcg	1-R 2-I 3-S
72	LVX	select_one	Levofloxacin	1-R 2-I 3-S
73	MNO	select_one	Minocycline	1-R 2-I 3-S
74	RIF	select_one	Rifampicin	1-R 2-I 3-S
75	TAZ	select_one	Tazobactam	1-R 2-I 3-S

76	TOB	select_one	TOB	1-R 2-I 3-S

77	COT	select_one	COT	1-R 2-I 3-S
78	LNZ	select_one	Linezolid	1-R 2-I 3-S
79	COL	select_one	Colistin	1-R 2-I 3-S
80	STR	select_one	STR	1-R 2-I 3-S
81	SPT	select_one	SPT	1-R 2-I 3-S
82	TCG	select_one	TCG	1-R 2-I 3-S
				1-R 2-I

83	FOX_ND10	select_one	Cefoxitin	3-S
84	NOR_ND5	select_one	NOR_ND5	1-R 2-I 3-S

Appendix B: Data Abstraction Checklist

1. Project Setup

Item	Status
Define the ML task (binary classification: MDR = Yes/No)	✓
Define MDR (e.g., resistance to ≥3 antibiotic classes)	✓
Secure ethics approvals / data sharing agreements	✓
Record all hospital sources (9 RRHs, 1 tertiary hospital)	✓

---

2. Data Sources and Access

Item	Status
Demographic data accessed (age, sex, region, etc.)	✓
Clinical data accessed (hospital stay, comorbidities, prior antibiotics)	✓
Microbiology data accessed (specimen type, isolate, antibiogram)	✓
Data source dates and versions recorded	✓

---

3. Variable Abstra ction

Category	Variables	Status
Demographic	Age, Sex, District, Facility	✓
Clinical	Admission date, Previous antibiotic use, Diagnosis	✓
Microbiological	Specimen type, *E. coli* isolate, AST results	✓
Outcome	MDR label (binary: 1/0)	✓
Facility-level	Hospital ID, Facility type (RRH/Tertiary)	✓

---

4. Data Cleaning

Item	Status
Resolve duplicates (especially in AST results)	✓

Handle missing values (drop, impute, or categorize)	✓
Harmonize antibiotic names and result formats (e.g., S/I/R)	✓
Standardize column names and formats	✓

5. Label Creation

Item	Status
Define MDR based on standard criteria (e.g., WHO)	✓
Create binary outcome variable (1 = MDR, 0 = Non-MDR)	✓
Check for class imbalance	✓

6. Dataset Integration

Item	Status
Merge demographic, clinical, and lab data using Patient or Sample ID	✓
Document merge type (e.g., inner join)	✓
Validate merge (e.g., % matched, unmatched samples)	✓

7. Final Dataset Review

Item	Status
Confirm variable types (categorical, numeric)	✓
Generate summary statistics (age range, sex ratio, etc.)	✓
Store final dataset securely (with date and version)	✓

8. Documentation

Item	Status
Create and maintain data dictionary	✓
Log all data cleaning and merge steps	✓
Note hospital-specific biases or peculiarities	✓