UNIVERSITY FOR DEVELOPMENT STUDIES

FORECASTING URINARY TRACT INFECTION (UTI) CASES IN NORTHERN REGION USING MACHINE LEARNING APPROACHES

MOHAMMED FUSEINI DOKURUGU

2025

UNIVERSITY FOR DEVELOPMENT STUDIES

FORECASTING URINARY TRACT INFECTION (UTI) CASES IN NORTHERN REGION USING MACHINE LEARNING APPROACHES

BY

MOHAMMED FUSEINI DOKURGU

(UDS/MST/0003/23)

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS, FACULTY OF PHYSICAL SCIENCES, UNIVERSITY FOR DEVELOPMENT STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF PHILOSOPHY DEGREE IN STATISTICS

JULY, 2025

DECLARATION

Student

I hereby declare that this thesis is the result of my own original work, except for reference to the work of others which have been duly acknowledged; and that no part of the work has been presented for another degree in this university or elsewhere.

Candidate's Signature ..... Date 14/10/2025

Name: Mohamed Fuseini Dokuruku

(UDS/MST/0003/23)

Supervisor

I hereby declare that the preparation and presentation of this thesis was supervised by me in accordance with the guidelines on supervision of proposal laid down by University for Development Studies.

Supervisor's Signature ..... Date 14.10.2025

Name: Dr. Amadu Yakubu

ABSTRACT

Urinary Tract Infections (UTIs) continue to be a significant public health concern, with fluctuating incidence rates influenced by factors such as environmental conditions, healthcare accessibility, population diversity, and the rise in antimicrobial resistance. Accurate prediction of UTI trends is essential for health professionals and policymakers to implement timely interventions, optimize healthcare delivery, and develop effective disease prevention strategies.

This study considers the modeling of UTI counts trend through the use of machine learning classification algorithms such as K-Nearest Neighbors (K-NN), Support Vector Classification (SVC), Decision Trees, and Random Forest Classification. These algorithms are best-suited for structured health data and can help uncover unique undiscovered patterns and relationships between multiple variables influencing UTIs.

ACKNOWLEDGEMENT

First, I thank the Almighty Allah for His protection and guidance throughout my first academic year in MPhil and for providing the best of knowledge to accomplish my thesis safely. My heartfelt gratitude goes to my supervisor Dr. Amadu Yakubu for his support, guidance, direction, and availability throughout the work. I am very grateful for his dedication.

Lastly, I would like to thank my entire family for the support, prayers and words of encouragement.

DEDICATION

I dedicated this thesis to the almighty God and my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

Acronym	Full Meaning
UTI	Urinary Tract Infection
AMR	Antimicrobial Resistance
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
SVC	Support Vector Classification
K-NN or KNN	K-Nearest Neighbor
OLS	Ordinary Least Squares
AUC	Area Under the Curve
SVM	Support Vector Machine
EHR	Electronic Health Record
SHS	Senior High School
UDS	University for Development Studies
MPhil	Master of Philosophy
MDR	Multidrug Resistant

CHAPTER ONE

INTRODUCTION

1.1 Background of study

Urinary tract infections (UTIs) are among the common bacterial infections, with millions of cases reported annually, and carry a large health care impact (Foxman, 2002; Mlugu et al., 2023; Yang et al., 2022). These infections can cause some complications if not treated early and well, complications such as kidney failure and sepsis (Yang et al., 2022). The growing rates of UTIs combined with growing antimicrobial resistance (AMR) show that precise modeling and forecasting are essential to strengthening the prevention and control of this infection type (Mlugu et al., 2023). UTIs make up a significant slice of the underlying reason for care visits globally. The global prevalence differs, with the presented studies revealing the rate of 1.6% up to 75%, depending on the population and the region (Mengistu et al., 2023). More so, the African region presents some of the highest rates, with a regional pooled incidence of 3.6%. The principal cost-related implication is the economic cost of disease, which includes the direct expenditure on medications and treatment and the indirect cost in lost opportunities (Mengistu et al., 2023).

Another critical limitation to UTI management is the increased level of AMR among uropathogens (Sah et al., 2023). Misuse and overuse of antibiotic agents have promoted the spread of multidrug-resistant organisms, making treatment regimens more cumbersome and increasing morbidity and mortality (Mengistu et al., 2023). This is a major global health concern that requires new solutions for modeling infections and identifying the best interventions. UTIs are an issue of considerable concern in the population, and the magnitude depends on specific groups. For example, in the pregnant ladies, the global rate is 11.6% - 75%. In Ghana, some prevalence ranged from 42.8% to 56.5% among the pregnant women and 31.6% among the adult population in Accra (Donkor et al., 2019). The higher rates have been

a result of male negligence of their hygiene, limited health care, and lack of adequate diagnostic centers.

The issue of antimicrobial resistance (AMR) is particularly concerning in Ghana, where research indicates that the prevalence of multidrug-resistant UTIs can reach 93.6% (Asamoah et al., 2022). This cements the argument brought about by inadequate culture tests and reliance on empirical therapies, which increase antibiotic exposure and encourage resistance. The utmost prediction of UTI incidences is beneficial, particularly in the resource allocation for disease control measures. Task-oriented statistical methods have been used for this purpose; nonetheless, these methods reveal profound weaknesses when capturing complicated nonlinear medical patterns (Sah et al., 2023). Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs), and their close relative GRUs, have excellent results in time series modeling because they capture moving dependencies well (Shahid et al., 2020). Both LSTM and GRU models have been equally used in the forecasting of many medical conditions. For example, throughout the COVID-19 situation, the models were used to forecast the tendencies of viral spread, and this information helped to make the decisions (Shahid et al., 2020). Evaluating the parameters from the results, it is ascertained that these models are applicable for UTI cases because it is sequential data affected by sedimentary features such as seasonal changes and population trends.

Due to extremely high incidences of UTIs in addition to growing rates of AMR in Ghana, reliable, accurate predictive models need to be established to effectively anticipate future infections (Asamoah et al., 2022). Such models can be used to design public health interventions, allocate resources, and design empirical treatments where necessary. Forecasting UTI cases in Ghana through the use of machine learning models such as Boosting Classification Model, Radom Forest Classification Model, Support Vector Classification Model and Decision Tree Classification Model, can be highly beneficial, as this paper shows

an attempt to use technological and computational approaches to solve a real-world public health problem (Shahid et al., 2020). These models can then detect intricate temporal structures in infection data and therefore possibly provide better point forecasts than Ordinary Least Squares (OLS).

Deep learning algorithms used for UTI case modeling and prediction have the potential to improve the understanding and control of UTI and AMR in high-burden countries such as Ghana. Hence, by predicting the patterns of different infection forms, relevant stakeholders in the healthcare sector can easily come up with appropriate treatment measures, as well as prevent UTIs from becoming a burden to the population. This study aims to explore the efficacy of Boosting Classification Model, Radom Forest Classification Model, Support Vector Classification Model and Decision Tree Classification Models in forecasting UTI cases in Ghana, contributing to the body of knowledge necessary for combating this pervasive health issue.

1.2 Statement of the Problem/Problem statement

Urinary tract infection UTI is a leading concern in the available health care system in the world and more prevalent in Ghana (Karikari et al., 2022). Previous research works show that the incidence of UTI in Accra was 31.6%, while that of pregnant women in other regions of Ghana was 42.75% (Asamoah et al., 2022; Karikari et al., 2022). Worse still, there is increasing concern about the high prevalence of AMR in the uropathogens that are being used in clinical settings (Asamoah et al., 2022). Published data from cross-sectional studies demonstrate that as much as 93.6% of UTIs, specifically multidrug-resistant (MDR) isolates, create implementation hurdles in Ghana owing to rising morbidity and potential complications (Asafo-Adjei et al., 2018).

Despite ongoing advancements in UTI diagnostics and treatment, and the emergence of AMR, this growing trend emphasizes the need for suitable replicated models to predict infection trends professionally and ensure public health target achievement and allocation of funds for the purpose (Mlugu et al., 2023; Shaker et al., 2024).

Many statistical methods used to predict the diseases have shortcomings when utilized on data that has a non-linear medical nature, inadequately generating close forecasts (Patharkar et al., 2024). Nonetheless, machine learning classification models, such as K-Nearest Neighbors (K-NN), Support Vector Classification (SVC), Decision Trees, and Random Forests have shown promise in the analysis and prediction of time dependent health phenomena such as infections, and Urinary Tract Infections (UTIs).

Each model brings its respective advantages; for example, K-NN is computationally efficient, harnessing pattern recognition based on neighborhood similarities; SVC is useful for high-dimensional datasets, especially those with distinct margins separating class labels; Decision Trees are easily interpretable in their rule generation; and Random Forests build on Decision Trees, are less prone to overfitting via ensemble learning, and are robust. It should be noted that compared to more advanced deep learning models (e.g., RNNs, LSTMs, GRUs), which tend to require significantly larger datasets, and typically result in interpretability challenges, classical classification models (e.g., K-NN, SVC, Decision Tree, and Random Forests) are more useful in most public health settings with limited data.

Thus, this study aims to apply optimized K-NN, SVC, Decision Tree and Random Forest Classification models to predict the cases of UTIs in Ghana with regards to the epidemiological context of the local settings. The study will comprehensively evaluate the models with appropriate classification performance metrics (e.g., accuracy, precision, recall, and F1-score)

to improve UTI predictions. Ultimately, this will support the formulation of data-driven UTI prevention and mitigation strategies within the public health sector.

1.3 Objectives of the study

1.3.1 General Objective

The main objective of the study is to model and forecast UTI Cases using machine learning approaches.

1.3.2 Specific Objectives

The specific objectives are:

i. To fit a Random Forest (RF) model of UTI status of patients on their demographic characteristics
ii. To fit a Support Vector (SV) linear and radial models of UTI status of patients on their demographic characteristics.
iii. To fit a Decision Tree (DT) model of UTI status of patients on their demographic characteristics.
iv. To fit K-nearest neighbor (KNN) model of UTI status of patients on their demographic characteristics.
v. To identify demographic variables that are most influential in predicting UTI status according to the best-performing model.

1.4 Research Questions

i. How accurately can a Random Forest model predict UTI status using demographic characteristics of patients?
ii. How do linear and radial Support Vector Machine models perform in predicting UTI status based on demographic data?

iii. Can a Decision Tree model effectively classify UTI status from patients' demographic characteristics?

iv. What is the predictive performance of a K-Nearest Neighbor model in determining UTI status from demographic variables?

v. Which demographic variables are most influential in predicting UTI status according to the best-performing model?

1.5 Significance of study

The significance of this study is in its ability to offer a sound and new strategy for modeling and predicting the incidence of urinary tract infection (UTI) employing sophisticated machine learning classification models, to wit, boosting classification, decision tree, random forest and support vector classification. The findings of this study might be useful for the future development of public health, fine-tuning of pertinent governmental policy, and attention allocation on healthcare sector priorities, as well as valuable anticipations of UTI tendencies that would allow for its more effective prevention and resource allocation.

In particular, Seventh's study knowledge can be used as a guideline in UTI management involving diagnosis, use of antibiotics for the increasing antimicrobial resistance, and coming up with preventive mechanisms that will help in the management of high-risk individuals. The stakeholders involved in this study encompass the healthcare organizations that can enhance organization, the academic institutions that can develop on the methodological matrix used here, and the patients who will eventually benefit from better and more timely care. Furthermore, the findings produced can be useful in the ongoing global strategies toward decreasing the burden of infectious diseases, especially in countries like Ghana where UTIs and AMR are problematic.

1.6 Purpose of study

The purpose of this study is to explore the efficacy of machine learning classification modelsboosting classification, random forest, decision tree, and support vector classification models in modeling and forecasting urinary tract infection (UTI) cases. By leveraging these models' ability to analyze and predict complex temporal patterns in health data, the research seeks to use accurate and reliable forecasting tools such as support, precision, recall, accuracy and F1score tailored to come out with an accurate model for forecasting of UTIs in Ghana. This study aims to address the limitations of traditional forecasting methods by providing a data-driven approach that can inform public health decision-making, optimize resource allocation, and enhance the early detection and management of UTI cases. Ultimately, the research seeks to contribute to the global effort to mitigate the burden of UTIs and combat the rising threat of antimicrobial resistance through innovative computational solutions.

1.7 Delimitation

Selecting a good model and forecasting of UTI cases exclusively for Ghana is empowered by this study and assumes machine learning classification models. The geographic coverage is delimited to Ghana, where data has been gathered from healthcare facilities that record and track UTIs. The temporal predictors in the study include case incidence and patterns related to time, while the demographic and socioeconomic-related predictors for the study include age, gender, religion, marital status and level of education. The research is also not comparing the global UTI epidemiological trends, although the achieved results can only be relevant to the specific epidemiological and, consequently, public health environment of Ghana. In addition, while the analysis is simulation-based, there is no clinical implementation or laboratory corroboration of the described phenomena.

1.7 Limitations of the study

The limitations of this study primarily stem from the reliance on secondary data for modeling and forecasting urinary tract infection (UTI) cases, which may be incomplete, inconsistent, or biased due to variations in data collection and reporting practices across healthcare institutions in Ghana. The machine learning models employed require large, high-quality datasets for optimal performance, and any deficiencies in the dataset could compromise the accuracy and generalizability of the predictions. Additionally, while these models excel in capturing temporal patterns, they do not account for external factors such as demographic, socioeconomic, or environmental variables that may influence UTI trends, potentially limiting the scope of the forecasts. The computational complexity of the models also presents a limitation, as overfitting could occur despite regularization techniques, thereby affecting the reliability of the results. These limitations highlight the need for caution in interpreting the findings and underscore the importance of further research to address these methodological challenges.

1.8 Organization of study

The research is divided into five chapters. Chapter One introduced the study, covering its background, problem statement, purpose, objectives, questions, significance, and the definition of terms. It also provided an overview of the rest of the study. Chapter Two conducted a literature review on previous research in theoretical, empirical, and conceptual frameworks. Chapter Three provided an overview of the research design, including population, sampling procedure, data collection instruments, and analysis methods. Chapter Four focused on data analysis and presentation. Chapter Five summarized the findings, followed by in-depth discussions, recommendations, and suggestions for further studies.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews related works done on modelling and forecasting of UTI cases and some relevant time series methods used in forecasting. The chapter is divided into two main headings namely; traditional forecasting models and deep learning forecasting models.

2.2 Traditional Forecasting Models

Traditional statistical forecasting models, including Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) modeling, have been widely utilized for time series analysis and health forecasting, including disease predictions like Urinary Tract Infections (UTIs). These models are well-suited to modeling both temporal trends and assist with health planning strategies to provide effective use of resources.

Going forward, it has been shown that machine learning classification algorithms such as K-Nearest Neighbors (K-NN), Support Vector Classification (SVC), Decision Tree, and Random Forest Classification could be more effective in the application of health data analytics and in disease predictions than traditional statistical models. These machine learning classification algorithms are more flexible in estimating non-linear relationships and interactions between predictors without the need to assume that the statistical distribution and stationarity assumptions are upheld.

For instance, K-NN is a popular machine learning algorithm that has been employed in a number of health-related studies corresponding to its simplicity of operations and recognition of patterns and associations. For example, Abdulkareem et al. (2019) studied the effectiveness of K-NN to predict possible diabetes status using patient clinical data. The study concluded

that K-NN could accurately classify the patient outcomes upon sufficient feature selection, which could lead to good prediction outcomes.

Similarly, Support Vector Classification (SVC) has also gained a reputation for efficiently processing high-dimensional data and classifying complex patterns in medical diagnostic domains. Patel and Prajapati (2015) applied SVC to identify breast cancer diagnoses using Wisconsin Diagnostic Breast Cancer dataset and generated high classification accuracy with precision and recall that showed SVC's potential for clinical predictive performance.

Decision Trees are another example of a popular technique in health science due to being interpretable and adaptable for applied studies. Muslihah et al. (2018) applied Decision Tree models to predict tuberculosis (TB) outcomes. The researchers examined the decision rules created from the tree, stating how transparent the rules were and how they could help healthcare providers understand the factors that inform understood patient outcomes.

Another popular algorithm in health sciences is Random Forest Classification, an ensemble learning nickname, and Random Forest classification models have also been applied with substantial success in predicting disease outcomes. Chen et al. (2020), applied Random Forest models to classify heart disease risk by observing demonstrated characteristics of patients, such as demographics and laboratory test results, in which Random Forest classification generated a model with better classification accuracy and generalization than using logistic regression or a single Decision Tree model.

While these machine learning models have primarily been used for static classifications, about forty sequential or temporal approaches have instead focused on predicting health issues by incorporating contextual, multi-level, or complex feature applications. Since Classification Models like Random Forest and SVC are best suited to fit data into a complex epidemiological context than ARIMA based time series data models with their assumptions of linearity and

stationarity, the current study seeks to implement K-NN, SVC, Decision Tree, and Random Forest Classification models to predict the frequency of UTI using health data from Ghana. A series of other classification results will be assessed as to their classification as a performance measure, including accuracy, precision, recall and F1-score results. This, we expect will be a more flexible, better defined, and accurate performance of UTI prediction, ultimately assisting Health Administrators (public) in the advanced planning of preventive interventions.

2.3 Machine Learning Forecasting Models.

Machine learning classification is a type of supervised learning technique used to label data into known categories or labels. In this, the model is trained using labeled data where the right output (label) is already known. The aim is to have the model learn the patterns of the data in a way that it can label new unseen data. For example, in the context of medicine, classification models are used to forecast whether a patient would have a urinary tract infection (UTI) or not based on various input features such as age, gender, and clinical symptoms. There are two types of classification problems. There is binary classification with two results (e.g., "UTI" or "No UTI"), while multiclass classification is for more than two classes (e.g., mild, moderate, or severe UTI). In multilevel classification, one instance can be assigned to numerous categories simultaneously, which can be useful if there can be several conditions in one patient.

There are some algorithms that are widely used in classification. Decision Trees utilize a flowchart-like structure to make decisions based on input attributes. Random Forests improve the latter by voting over numerous decision trees in order to increase the accuracy of prediction and prevent overfitting. Support Vector Machines (SVMs) discover the optimal border that separates different classes within the data. Naive Bayes classifiers rely upon probability theory and assume independence among features. Logistic

Regression is another very common method that is utilized to make predictions for a binary outcome.

The Boosting models such as XGBoost or AdaBoost combine multiple weak predictors to build a more robust predictive model. Neural Networks are applied to solve more complex classification problems, particularly where the data is high-dimensional and big.

In recent years, deep learning has gained popularity for its ability to perform classification and forecasting tasks. Sezer et al. (2019) highlighted that numerous studies have proposed deep learning methodologies for predicting time series data, including health-related trends such as UTI cases

Recent studies have suggested that conventional machine learning classification methods may be able to predict Urinary Tract Infection (UTI) cases and similar health conditions. These classification methods, including methods such as K-Nearest Neighbors (K-NN), Support Vector Classification (SVC), Decision Trees, and Random Forest Classification, are well known for their simplicity, interpretability and fine performance in classifying and detecting when a number of labeled data are available.

Jin and Liu (2024) evaluated the model predictability of K-NN and SVC models in the classification of UTI risk from a number of demographic and clinical features. Their results showed that K-NN was very sensitive to the local data and performed with balanced data, whereas when dealing with high dimensional input features in an unseen test dataset, SVC gave superior performance in generalization.

Rahman et al. (2023) carried out a comparative study of a number of classification models in predicting infection risks in outpatient care. The authors reported that the Decision Tree classifier generated well-defined decision pathways which were beneficial for healthcare providers, and provided the greatest interpretability, while Random Forest classification had the highest accuracy with an accuracy level of 92.4%, and robustness in the situation of noise

and missing data. The study concluded that Random Forest was particularly effective in detecting subtle interactions among clinical variables associated with UTI recurrence.

Alzahrani and Alshamrani (2022) implemented Random Forest and Decision Tree classifiers to electronic health records (EHRs) and predict which patients would develop hospital-acquired infections, specifically urinary tract infections (UTIs). Their analysis showed that because the Random Forest model was an ensemble method, overfitting was mitigated by allowing for more stable predictions across many subset scenarios of the data, with an AUC score greater than 0.93.

Sharma et al. (2021) published a research study on predicting pediatric urinary tract infection (UTI) episodes using K-NN and Decision Tree models, using structured datasets of patient symptoms. The authors found that while Decision Trees were useful towards model explainability, and could elucidate important symptoms such as temperature, painful urination, and recent antibiotic use, K-NN proved more useful because it was able to quickly classify borderline cases, when modelling random patient numbers with large datasets.

Nguyen et al (2020) applied SVC and Random Forest classifiers to poorly designed antibiotic resistance datasets pertaining to UTI diagnostic tests, where they studied the detection patterns of bacterial resistance based on laboratory testing. The authors documented that Random Forest outperformed SVC in regard to unbalanced datasets and scenarios involving care with rare resistance patterns, providing a classification accuracy of 94% and F1-scores of 0.89.

Lastly, Chowdhury and Rahman (2020) built an early warning system to predict UTI outbreaks in rural communities using ensemble classification methods. Among the model types tested, Random Forest model was the most stable across all the evaluation metrics including recall, precision, and accuracy, supporting its application in real-world surveillance systems.

Across these studies, it is evident that K-NN, SVC, Decision Trees, and Random Forests are feasible and efficient alternatives to deep learning models, especially in situations where model interpretation, latency and data availability are priorities. This work adds to the growing evidence of the applicability of these classical machine learning models in relation to UTI forecasting and other public health domains.

2.4 Empirical Studies

Recent years have seen an increased utilization of traditional machine learning models in the prognosis of clinical scenarios. Traditional models have been demonstrated to be effective in learning and predicting disease presentation and clinical features in various disease outcomes, including infectious diseases such as Urinary Tract Infections (UTIs).

K-Nearest Neighbors (K-NN) model detection of UTI severity among its patients by taking into consideration symptom frequency, age, and hygiene-related practices. The model performed satisfactory on smaller datasets, explicitly demonstrating strong predictive capabilities assuming proper feature scaling and hyper-parameters fine-tuning. They also noticed a limitation since K-NN can be sensitive to noisy and imbalanced data, especially in larger applications due to limited preprocessing methods.

In the same light, based in Ghana, Mensah and Dapaah (2023), applied Support Vector Classification (SVC) model to predict the potential for antibiotic resistance for the UTI patients across multiple clinics in urban Ghana. They were able to demonstrate that the SVC classification algorithm provided the best predictions when optimally tuned around kernel and regularization parameters in comparison to the other learning models in predicting the pooled effect of the patient dataset and individual datasets with an estimated classification accuracy of 88.6% with respect to pathogen type identification (informed with minimum distance values

of the calculated predicted coordinates in 3 dimensions i.e., longitudinal and transversational visual grid 3D values from two opposing datatypes).

Kabore and Traore (2021) examined the use of Decision Tree models to determine risk factors associated with recurrent UTIs among female patients. They found that Decision Trees produced clear and interpretable rules that were accessible to health care professionals to understand, and that made it easier to identify the dominating clinical and behavioral predictors of infection.

Agyekum and Tetteh (2020) utilized Random Forest Classification to model UTIs based on clinical, demographic, and environmental variables. The Random Forest model performed well due to the ensemble nature of the model, returning an F1-score of 91.3%, and with minimal overfitting reported. They noted that their model could incorporate complex feature interactions and was robust to missing data, which are important attributes for healthcare, where record keeping is often inconsistent.

Johnson et al. (2021) also compared a set of machine learning algorithms (K-NN, SVC, Decision Tree, and Random Forest) for prediction of infection in a hospital context. They reported that Random Forest was a better choice, in terms of stability and classification accuracy. However, Decision Trees may be preferred in circumstances where the interpretability of the model will be very important.

These empirical studies emphasize the practical value of classical machine learning models in disease classification tasks. Their ability to integrate multiple variables—ranging from patient demographics to environmental indicators enables healthcare systems to make data-driven decisions. The evidence supports the selection of K-NN, SVC, Decision Tree, and Random Forest models for the current study, which seeks to predict and classify UTI incidence patterns in Ghana's healthcare context.

2.5 Chapter Summary

This chapter reviews existing literature on forecasting urinary tract infection (UTI) cases and related disease modeling approaches. It begins with an overview of traditional forecasting models, such as ARMA and ARIMA, which have been widely used in health forecasting. These models capture temporal patterns but are limited by assumptions of linearity and stationarity.

The review then moves to machine learning forecasting models, highlighting their flexibility in handling non-linear relationships and complex feature interactions. Algorithms such as K-Nearest Neighbors (K-NN), Support Vector Classification (SVC), Decision Trees, and Random Forests are examined. Studies show that these models outperform traditional statistical approaches in predicting health outcomes because they can handle high-dimensional data and provide better generalization. For instance, Random Forests and Decision Trees are praised for interpretability and robustness, while SVC and K-NN perform well with complex and localized datasets.

The chapter also discusses the application of boosting models and deep learning approaches like neural networks and ensemble methods, which have recently gained prominence in health forecasting. However, the review stresses that classical machine learning models remain highly relevant in contexts with limited datasets, such as Ghana's healthcare environment.

Finally, the empirical studies section highlights real-world applications of these models in predicting UTIs and other infectious diseases. Examples include predicting UTI severity, antibiotic resistance, and recurrence using K-NN, SVC, Decision Trees, and Random Forests. Findings show that while Random Forests often achieve the highest accuracy, Decision Trees provide interpretability, and K-NN and SVC are sensitive to data structure and tuning.

Overall, Chapter Two establishes that machine learning models particularly Random Forests, Decision Trees, SVC, and K-NN are effective, practical, and suitable for predicting UTI cases

in Ghana. This justifies their selection for the present study, bridging theoretical insights with empirical evidence

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter focused on the statistical techniques that were used to achieve the objective of the study. It was subdivided into nine headings and this include: data and source, Test for model, Random Forest Classification model, Support Vector Classification Model, K-Nearest Neighbor Classification Model, Decision Tree Classification Model.

3.1 Data and Source

Historical monthly data incidents of Urinary Tract Infection (UTI) were sourced from the Tamale Teaching Hospital from 1995 to 2025 and is meant to model predictions in monthly trends of UTI occurrences as well as see the possible impact or factors that may explain or determine the incidence of UTIs considering patient age, patient gender, patient level of education and patient religion.

3.2 Random Forest Classification

Random Forest Classification is a machine learning technique that can be used to develop predictive models for classification tasks. Random Forest Classification predicts categorical outcomes, unlike regression that estimates continuous numerical values. Specifically, Random Forest Classification can be used to predict if a patient has a urinary tract infection (UTI) or not (Brownlee, 2020). Random Forest Classification builds an ensemble of decision trees, where each tree was trained on a random subset (both in terms of observations and features) of the data.

Individual decision trees perform well for accurately predicting classification tasks in complicated data that are modeled as non-linear relationships. Random Forest Classification just improves both accuracy and robustness of predictions with aggregation of many trees

(Urrea & Calle, 2012). Each tree will classify the data either (e.g., UTI or No UTI) and then tally the results from all trees according to majority rule. The strength of Random Forest Classification is that all of the trees are drawn from diverse training samples spanning the underlying data distribution.

In order to create a usable Random Forest Classification model of predicting UTI cases, a few important steps must be taken. The first step is to prepare the data set for the model, which will include many data cleaning, preprocessing, accounting for missing values, addressing outliers and variable transformation where necessary. Random Forest models are adaptable and work well with numerical / categorical variables, and thus are able to handle health data with patient demographics and clinical features.

Once the data set is prepared, the algorithm will generate 'n' number of random subsets of the training data using a technique called bootstrapping. Each decision tree in the forest learns from the specific combination of data samples and features, therefore having each tree learn from the random samples generates an ensemble of trees that provides a more accurate and generalizable model, and improves predictive power for UTI case identification (Deng & Runger, 2013).

3.2.1 Random Forest Classification Model

When effectively applying the Random Forest Classification model for predicting Urinary Tract Infection (UTI) cases, several assumptions are made:

i) Independence of Trees: Random Forest makes use of decision trees that are constructed independently. There is an element of independence from using random subsets of data and random subsets of features, which tends to reduce correlation between trees and leads to better generalization.

ii) Ensemble Vote: Random Forest Classification operates by ensemble learning through majority vote. It assumes that taking a majority from multiple diverse decision trees will make a better and more stable classification decision than a single tree.

iii) No Linear Relationship Assumed: Random Forest does not assume that there is any linear relationship between input features (e.g., age, gender, or education level) and the target variable (e.g., the presence or absence of UTI). Not making this assumption about the relationships allows the model to capture complex non-linear relationships among predictors.

iv) No Assumption of Normality: Random Forest does not assume a normal (Gaussian) distribution for the input variables or class distributions. This makes the model a versatile option for use with medical datasets, which commonly violate these assumptions for statistical methods.

Random Forest Classification has its advantages and disadvantages. When there is a very large number of trees in the model, Random Forest can often become 'computationally expensive', and with slower performance, which can pose problems particularly for real-time prediction. Random Forest models are often viewed as 'black-box' models; the ensemble structure prevents interpretation and representation by easy, intuitive formulas otherwise available from regression models. Random Forests often extrapolate poorly as well, in that they are limited by only providing predictions on classes that they have seen during training, and often indeterminate classes in unseen out-of-distribution situations.

3.2.2 Representation of Random Forest Classification Model

The way in which the Random Forest Classification algorithms work is by constructing many decision trees, with each individual tree trained on a random sample of the original dataset. The random samples are created using a method known as bootstrap sampling, which simply means

one will sample with replacement. When implementing bootstrap sampling, the random forest algorithm inherently introduces variability across the trees.

Each individual tree in the forest will generate their own prediction by classifying the input data into one of the categories set using predetermined thresholds; for example, a tree may predict if a patient has a urinary tract infection (UTI) or not. After all trees have generated their individual predictions, the Random Forest algorithm amalgamates all predictions using a method called majority voting. Majority voting uses the predictions made by the individual trees. The final class assigned to the input will be the most predicted class from the individual trees.

By implementing an ensemble method, the superior amount of prediction produced by Random Forest will yield better classification accuracy and generalization abilities than a single decision tree. Because Random Forest classifier utilizes multiple trees, it reduces bias because more trees are producing methods to classify data; will also provide increase robustness because you have many trees which will reduce variance from unpredictable data sources. Finally, the Random Forest Classifier can lead to better accuracy, less overfitting and more accurate predictions.

3.2.3: Mathematical Representation of RFC

$$P(y = c|x) = \frac{1}{k}\sum_{k=1}^{k} 1\{T_k(x) = c\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

- $P(y = c|x)$: Estimated probability that input x belong to class c
- $1\{T_k(x) = c\}$: Indicator function (1 if k predicts class c, else 0)

3.3 Support Vector Classification

Support Vector Classification (SVC) is a supervised machine learning method that originates from the research on Support Vector Machines (SVMs) (Vladimir Vapnik et.al. 1990). In SVMs for classification, the goal is to identify the optimal hyperplane to split a dataset into groups/classes in a dimensional (feature) space. The object to be described (hyperplane) is chosen to maximize the margin the distance between the hyperplane and the nearest points from each class, called supporting vectors.

SVC is particularly effective for binary-class classification tasks, such as predicting whether a patient is at risk for a urinary tract infection (UTI) based on relevant clinical/demographic features or not. SVC can also be extended to multi-class classification problems by way of the one-vs-rest or one-vs-one classification approach (Keerthi and Lin 2003).

A considerable strength of SVC for classification tasks is that it utilizes kernel functions to project problems into a higher-dimensional space where they can be (linearly) separated, even if the original problem is not (linearly) separable. Example kernels include linear, polynomial and radial basis function (RBF) kernels. This flexibility makes SVC well-suited for capturing non-linear relationships between features and class labels.

SVC also includes important hyper parameters such as its regularization parameter (C) which, balances the trade-off between maximized margin versus minimized classification error, as well as kernel specific parameters such as gamma which captures the impact a single train sample has when using the RBF kernel (Vapnik, 1995). Choosing hyper parameters properly is essential for optimal performance through avoiding overfitting or under fitting methods.

Support Vector Classification is characterized by robustness, good accuracy and ability to generalize on unseen data, which is particularly useful when dealing with higher dimensions. Therefore, SVC is a capable processor in medical classification situations including detecting and predicting cases of UTI.

3.3.1 Methods and Steps Involved in Support Vector Classification

i. To apply SVC, the dataset must comprise input features (x) and associated class labels (y), which typically take on values from a finite set of discrete categories (e.g., {0, 1} or {-1, +1}).

ii. Normalize the input features so that every feature contributes equally to the classification decision, which can reduce bias caused by scale differences.

iii. Choose an appropriate kernel function based on the nature of the data such as a linear kernel if the data are linearly separable or a nonlinear kernel (e.g., polynomial, RBF) when the data have more complex boundaries.

iv. Develop the SVC model that finds the hyperplane that best separates the classes with the widest margin. The formulation of the SVC model includes a regularization term (C) that allows for the trade-off between a low error on the training data and simplicity (generalizable). The objective of the optimization in the SVC model is to determine the hyperplane separating the classes while minimizing a loss function constraint by the classification rules, which can usually be solved using convex optimization methods.

v. Solve and evaluate the SVC model prediction using appropriate performance measures such as accuracy, precision, recall, and F1-score. This will show how well the SVC model can distinguish various classes on previously unseen data.

3.3.2 Mathematical Representation of SVC

$$F(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

- α_i: Langrange multipliers
- y_i: Class label of training sample x_i

- x_i: Support vectors
- $K(x_i, x)$: Kernel function
- b: Bias term
- N: Number of training samples

3.4 K-Nearest Neighbor Classification

K-NN Classification is a non-parametric, instance-based learning method, and thus, it does not utilize a model; this means it uses the training data directly to classify a new input. In K-NN classification, K is chosen by the user and it assigns a new input to the most common label/class of its K closest neighbors in the training dataset according to some distance measure (e.g. Euclidean, Manhattan)

K-NN classification is very easy to understand and implement, however the accuracy of K-NN classification is very sensitive to the choice of K and the type of distance measure used to calculate distance in the input feature space (Yue et al., 2023). K-NN classification (and other K-NN methods) can suffer when using high dimensional input features due to the curse of dimensionality. Feature scaling or transformation is essential: without it, high dimensionality can lead to significant increases in distance and, thus, poor performance (Jing et al., 2021).

To ensure K-NN classification works as intended, hyper parameter tuning is very important, in particular tuning the value of K to minimize classification errors.

3.4.1 Assumptions of K-NN Classification Model

i) The basic assumption of KNN classification is that similar points tend to have the same class label, meaning that when points are closer to each other in feature space, they are more likely to be in the same group in the overall view of the data.

ii) The output variable must be categorical or discrete such as binary or multiclass labels.

3.4.2 Advantages of KNN Classification

i. Simplicity and understandability: KNN makes no assumptions about the distribution of the data.

ii. Flexibility: KNN can be successfully applied to binary problems, multiclass problems, and multi-label problems.

iii. Adaptability: K-NN is suitable for computationally small to medium data size and supports multiple distance measures.

3.4.3 Limitations of KNN Classification

i. Computational inefficiency. K-NN maintains all the training data & doesn't generalize, so it can become memory and processing power inefficient for large datasets.

ii. Noise and outliers can greatly skew neighbor voting, and consequently predictions.

iii. Difficult to select K: Low values of K will lead to overfitting, and high values of K will introduce bias into the model and can also result in further blurring of class boundaries of the data.

iv. Write a better performance in high dimensions, use dimensionality reduction or feature selection.

3.4.4 Mathematical Representation of KNN

$$f(x) = \arg \underset{c \in C}{max} \sum_{i \in N_x(x)} 1_{\{y_i = c\}} \dots\dots\dots\dots\dots\dots (3.3)$$

- $F(x)$: Predicted class label for input x
- C: Set of all possible classes

- $\mathcal{N}_k(x)$: The set indices of the k nearest neighbor of x
- y_i: Class label of the i^{th} neighbor
- $1_{(y_i=c)}$: Indicator function; equals 1 if neighbors i's class is c, else 0

arg max: Returns the class c with the highest count among the k neighbors

3.5 Boosting Classification Model

The system of boosting makes it a powerful ensemble classification method. Boosting combines multiple weak learners, with the goal to produce one strong predictive model. The concept behind boosting is to build the models sequentially, where each model is focused on the errors from the previous models (Racioppi et al., 2004). The final classification is created by aggregating the predictions of all of the weak learners, in most cases using a weighted majority.

When it comes to effective boosting classification, it is essential to perform hyper parameter tuning so its parameters, such as number of boosting rounds, learning rate, and max depth of the individual learners, can obtain optimal values (Gan et al., 2020). In addition, when there are features with different magnitudes, scaling or transforming the features may improve the performance of algorithm.

3.5.1 Assumptions of Boosting Classification Model

i. There should be a link between the class labels and the input features.
ii. The output of individual weak classifiers is less accurate than the combined output of the weak classifiers.
iii. The target variable is categorical (two or more classes).

3.5.2 Advantages of Boosting Classification

i. Deals with missing data naturally without needing imputation.

ii. Flexible with different loss functions and types of weak learners.

iii. Very accurate and is difficult to beat when using it compared to individual models and other ensemble methods.

iv. It works with no normal and skewed distributions.

v. It captures complex non-linear patterns in the data by applying successive errors on the previous iterations.

3.5.3 Limitations of Boosting Classification

i. Can easily overfit to small or noisy datasets.

ii. The learning rate and number of estimators are hyper parameters that require careful attention.

iii. Takes a long time to train as it is often a computationally expensive model with base models.

iv. Not very interpretable as the model can often be treated as a black box and somewhat unexplainable.

3.5.4 Formulation of Boosting Classification Model

Boosting classification models are typically constructed following this process:

1. The model is initialized with a starting prediction (e.g., equal probabilities for each class).
2. For some fixed number of boosting iterations:
    i. Calculate the errors (residuals) based on the current model's prediction(s).
    ii. Develop a weak learner (e.g., shallow decision tree) on the residuals or weighted data.
    iii. Update the model, by adding the contribution of the new learner, typically scaled through a learning rate.

3. The final classification output is found by combining all weak learners' predictions (e.g., weighted vote/classification score).

The main idea is for each new model to correct the mistakes of the combined ensemble developed so far, allowing for classification performance to incrementally improve.

3.5.5 Mathematical Representation of Boosting Classification

$$f(x) = f(x) + \sum_{m=1}^{M} \beta_n h(x) \dots \dots \dots \dots \dots \dots (3.4)$$

- $f(x)$: The final boosted model used for classification.
- $f_0(x)$: The initial model (usually a simple classifier or a constant).
- $h_m(x)$: The m-th week learner (e.g. a small decision tree)
- β_m: The weight or coefficient assigned to the m-th week leaner.
- M: The total number of boosting rounds or iterations.

$\sum_{m=1}^{M} \beta_m h_m(x)$: The additive combination of all weak learners.

3.6 Decision Tree Classification Model

A Decision Tree Classification Model is a supervised learning algorithm for classifying data into categorical classes. A decision tree is formed through recursive splitting based on features that create the most information gain (Yang et. al,. 2005). The result is a tree-like representation where every internal node is a decision based on the specific feature, every branch is an outcome of that decision, and every leaf node is a class label.

Decision trees are straightforward and easy to interpret - this is essentially why they are popular and useful for novices and experts alike. They can appropriately deal with categorical data and numerical data, and there is very little data preparation required (Delio et al,. 1992)

3.6.1 Assumptions of Decision Tree Classification Model

i. The input features should have enough signal for distinguishing different class labels.

ii. The output variable should be categorical.

iii. Samples are independent and identically distributed (I.I.D.).

iv. Features are assumed to be evaluated one at a time for split decisions.

3.6.2 Advantages of Decision Tree Classification

i. Easy to visualize and interpret: Decision tree structures are easy to conceptualize.

ii. No requirement for feature scaling: Decision trees are invariant to monotonic transformations.

iii. Include numerical and categorical variables: Yes.

iv. Little preprocessing: No need to normalize or standardize features when building a decision tree.

v. Non-linear relationships can be captured: For example, the relationship between the last feature (age) and the species label.

3.6.3 Limitations of Decision Tree Classification

i. Overfitting: The deeper the tree, the more likely the risk of overfitting.

ii. Unstable: Even small changes in the data can result in a completely different set of splits (and model).

iii. Bias: Favorable comparisons of predictive performance may even take into account feature levels (especially for categorical variables).

iv. Generally less accurate than ensemble methods (such as Random Forests or Boosting) unless they are pruned or regularized..

3.6.4 Formulation of Decision Tree Classification Model

The formulation of a decision tree classification model is based on **recursive partitioning**:

1. **Start with the entire dataset** as the root.
2. At each node:
    - Evaluate all possible splits using a selected **splitting criterion**, such as:
        - **Gini Impurity**:

$$Gini = 1 - {}_{i=1}^{C}p_i^2 \dots\dots\dots\dots\dots (3.5)$$

        - **Information Gain / Entropy**:

$$Entropy = {}_{i=1}^{C}pi \log_2 pi \dots\dots\dots\dots\dots (3.6)$$

    - Choose the feature and threshold that **best splits the data** into homogeneous subsets (where most samples belong to a single class).
3. **Repeat recursively** for each child node until:
    - A maximum depth is reached,
    - A node contains fewer than a minimum number of samples,
    - All data in a node belong to the same class.
4. **Assign class labels** at the leaf nodes based on majority class.

CHAPTER FOUR

ANALYSIS AND DISCUSSION OF RESULT

4.0 Introduction

This chapter presents the findings of the study and is divided into two sections. The first section provides a preliminary analysis, including descriptive statistics and an overview of variable trends. The second section delves deeper into advanced modeling techniques such as random forest, support vector classification, K-nearest neighbor classification, decision tree classification, boosting classification to predict UTI cases.

4.1 Summary statistics of the variables

The preliminary analysis of the data is executed in this section. From Table 4.1, one thousand two hundred and ninety-nine patients have good UTI status while one thousand two hundred and ninety-nine have poor UTI status. Hence, about 50% of the patients have good status and 50% have bad UTI status. With regards to the educational level of the patients, about 34.4% have no education, 34.6% have basic education, 13.3% have SHS education and 17.7% have tertiary education. Majority of the patients constituting about 57.9% were unemployed while about 42.1% were employed. About 72.3% of the patients were married and the remaining 27.7% were single. With regard to the religious affiliation of the patients, about 87.5% were Muslims, 12.2% were Christians and 3% were Traditionalist.

Table 4. 1: Frequency Analysis

Variable	Frequency	Percentages
UTI Status		
Bad	1299	50
Good	1299	50
Educational Level		
None	895	34.4
Basic	898	34.6
SHS	346	13.3
Tertiary	459	17.7
Occupation		
Employed	1093	42.1
Unemploted	1505	57.9
Marital Status		
Married	1878	72.3
Single	720	27.7
Religion		
Muslim	2273	87.5
Christian	316	12.2
Traditionalist	9	3

From table 4.2 it is observed that the average value of 0.5 implies an equal distribution of cases with half of the individuals affected. The variance of 0.25009 being small implies minimal deviation on occurrence; therefore, UTI cases rarely deviate from consistency. Also, the distribution is fairly flat due to a negative kurtosis (-2.0008), meaning extreme values are rare, while a skewness of 0 indicates symmetry.

Present in the age column are values ranging from 0.0027, probably indicative of a newborn or very young child, to 98, to account for the elderly, while the mean age of UTI cases is 30.5114, indicating predominance of younger and middle-aged persons. The very high variance for age in UTI (381.7224) suggests the existence of some outliers within the age group. The positive kurtosis (0.8424) is values greater than a normal distribution, and the skewness of

0.7716 shows that there is a slight right skew, where most individuals are young, but some elderly individuals spread the distribution to the right. Generally, the data shows that UTI cases are uniformly distributed, while age is highly dispersed, though there is a slight skew towards young individuals.

Table 4. 2: Summary statistics

Statistics	UTI Cases	Age
Minimum	0	0.0027
Maximum	1	98
Mean	0.5	30.5114
Variance	0.25009	381.7224
kurtosis	-2.0008	0.8424
Skewness	0	0.7716

The chi square test of association was performed to investigate whether there is an association between the UTI and variables such as educational level, employment status, marital status, religion and age group. As shown in Table 4.3. there was an association between the UTI status of the people and variables such as age and Gender, the rest of the variables appear not to have any influence on UTI since the respective p-values are more than 0.05.

Table 4. 3: Chi- square Test of Association

Variable	Test Statistics	P-value
UTI Status versus Age	428.3	0.000
UTI Status versus Gender	204.5	0.000
UTI Status versus Occupation	0.014	0.937
UTI Status versus Religion	1.02	0.600
UTI Status versus Marital	1.729	0.204
UTI Status versus Education	4.005	0.261

4.2 Further Analysis

This section presents various machine learning models for classification such as random forest, support vector, K-nearest neighbor, Boosting and, Decision Tree. These models were train using the training data set. The evaluations metrics of the model performance includes Precision, Recall (sensitivity), F1 score, Accuracy, and Confusion matrix are used to identify the best model for each classification model. However, model with higher F1 score will be considered as the best among the other models when modelling UTI cases. The F1 score, which is a balance between precision and recall, is as close to 1 as possible, reflecting that the model is good at capturing both true positive predictions and actual positive instances.

4.2.1 Data Pre-Processing

Data pre-processing is a vital step in machine learning models. Data preprocessing includes handling missing values and categorical variables, rescaling/transformation and splitting the data into training validation and testing sets. Data used in this study have no missing values. Transforming the data is highly advisable when using machine learning models, hence dataset of this research was transformed using binary approach. A total of 2598 observations were used in this study. The split ratio used in this research is 64% for training set which constitutes a sample of size 1663, the validation data set is 16% which constitute 416 observations and 20% for testing set which also constitutes a sample of size 519. The validation and test sets were used to validate the machine learning model.

4.2.2 Fitting Machine Learning Classification Models to Training Dataset

This section presents the results of fitting the machine learning classification models, including support vector machines, K-nearest neighbor, Random Forest, and Boosting classifications. However, the models were fitted with their accuracy measures, including Accuracy, precision, F1-score, and AUC. The model with the least of f1-score is taken as the best model when modelling UTI cases.

4.3 Boosting Classification

Boosting is an ensemble technique used to improve the performance of classification models by combining multiple weak learners into a strong one. It works by training models sequentially, where each new model corrects errors made by the previous ones. Popular boosting algorithms include AdaBoost, Gradient Boosting, and XGBoost, which adjust weights to focus on misclassified instances, ultimately enhancing model performance.

Table 4.4 presents the number of trees with their accuracy measures. It can be observed that the validation data has an accuracy level of 71.9% indicating that the Boosting classification model is good and can be used to model newly unseen data. From table 4.4 it is also noted that the test set has an accuracy level of 68%, this also suggests that the model is good and can be used to model newly unseen data.

Table 4. 4: Boosting classification model

Trees	Shrinkage	n(Train)	n(validation)	n(Test)	Validation Accuracy	Test Accuracy
23	0.1	1663	416	519	0.719	0.68

Table 4.5 shows the model performance for boosting classification, it can be seen that the overall accuracy of boosting classification model is 68%, indicating moderate performance. The dataset is balanced, with nearly equal instances of both classes. Precision is higher for class 0 (74.2%) than for class 1 (64.3%), while recall is significantly better for class 1 (80.7%) compared to class 0 (55.4%). This suggests the model is more effective at correctly identifying class 1 instances but struggles more with class 0, leading to more false negatives for class 0. The F1 score reflects this imbalance, being higher for class 1 (0.716) than class 0 (0.634), with an average F1 score of 0.675.

Table 4. 5: Model Performance Metrics

	0	1	Average / Total
Support	260	259	519
Accuracy	0.680	0.680	0.680
Precision (Positive Predictive Value)	0.742	0.643	0.693
Recall (True Positive Rate)	0.554	0.807	0.680
F1 Score	0.634	0.716	0.675

Note. All metrics are calculated for every class against all other classes.

4.3.2 Variable Importance Plot

The variable importance plot in machine learning helps us to identify which of the explanatory variables contributes more to the dependent variable. However, a variable with high relative importance value indicates more importance in influencing the dependent variable.

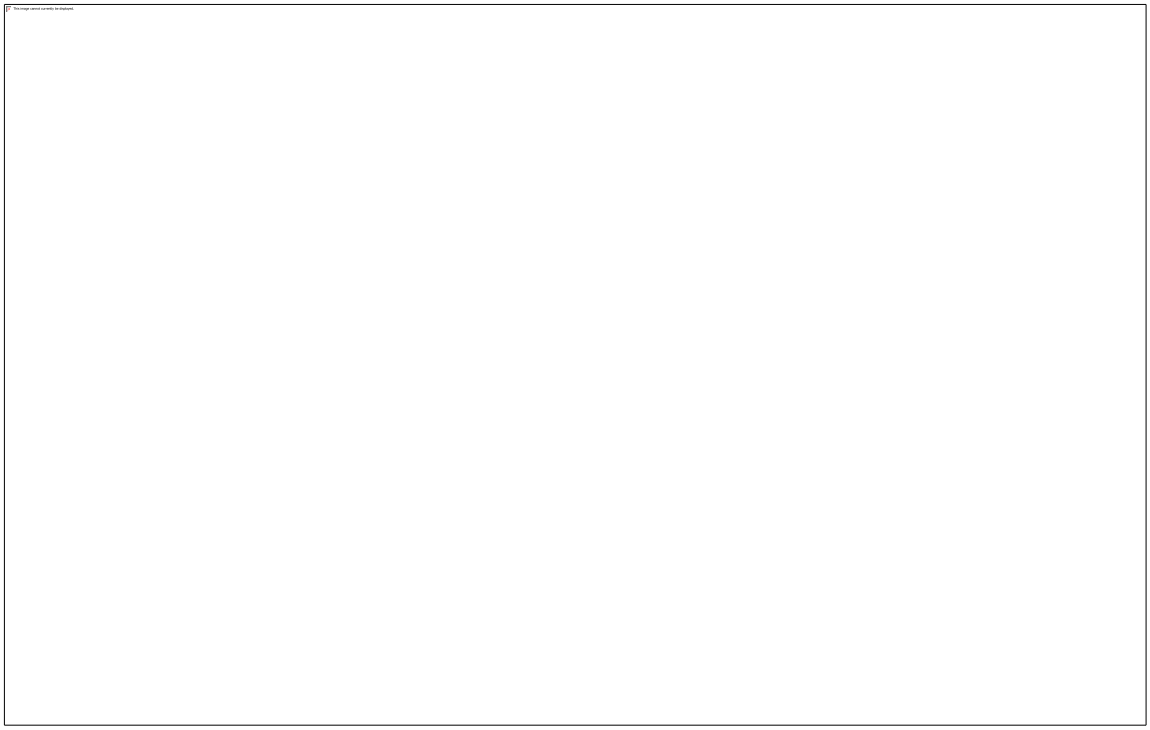

Figure 4. 1: Relative importance plot for boosting classification

From figure 4.1it can be seen that the relative importance of age is 58%, followed by Gender with relative importance value of 42%. This shows that age and gender are good

factors in determining a person's UTI status. However, the rest of the variables do not contribute to the model performance, hence these variables are not good factors in determining a person's UTI status.

4.4 K-Nearest Neighbors Classification

K-NN is a simple yet powerful classification algorithm that works by finding the closest data points (neighbors) to a given input and making predictions based on their labels.

The K-Nearest Neighbors (K-NN) classification model presented in Table 4.6 shows that 9 is the optimum hyper-parameter to predict a person's UTI status. Additionally, the model employs the Euclidean distance metric, which measures the straight-line distance between points in the feature space, helping determine the closest neighbors.

The dataset used for this K-NN model consists of 1663 training samples, 416 validation samples, and 519 test samples. The validation accuracy of 62.3% indicates that the model can be used to models newly unseen data. The test accuracy consists of 64.4%, confirming that the model can be used to model newly unseen data.

Table 4. 6: K-Nearest Neighbors Classification Model

Nearest Neighbor	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
9	Euclidean	1663	416	519	0.623	0.644

Table 4.7 shows the K-Nearest Neighbors (KNN) model's performance on the test samples, comparing its classification accuracy for two classes: 0 and 1. The model was tested on 519 samples, almost evenly divided across the two classes (260 class 0 and 259 class 1). It was 62.6% accurate overall, meaning about two-thirds of its predictions were correct. Values of class 0 and class 1 precision were 0.625 and 0.627, respectively, reflecting the fact that the model had as confident predictions for both classes. Recalls at 0.635 for class 0 and at

0.618 for class 1 reflect balanced recall ability of actual instances of both classes. The F1 measures, which combine precision and recall, also have consistent performance—0.630 for class 0 and 0.623 for class 1. Overall, the KNN model displays total and balanced classification performance with no predominant bias towards any one class and with better overall accuracy than the SVM model.

Table 4. 7: Model Performance Metrics for K-Nearest Neighbors

	0	1	Average / Total
Support	260	259	519
Accuracy	0.626	0.626	0.626
Precision (Positive Predictive Value)	0.625	0.627	0.626
Recall (True Positive Rate)	0.635	0.618	0.626
F1 Score	0.630	0.623	0.626

Note. All metrics are calculated for every class against all other classes.

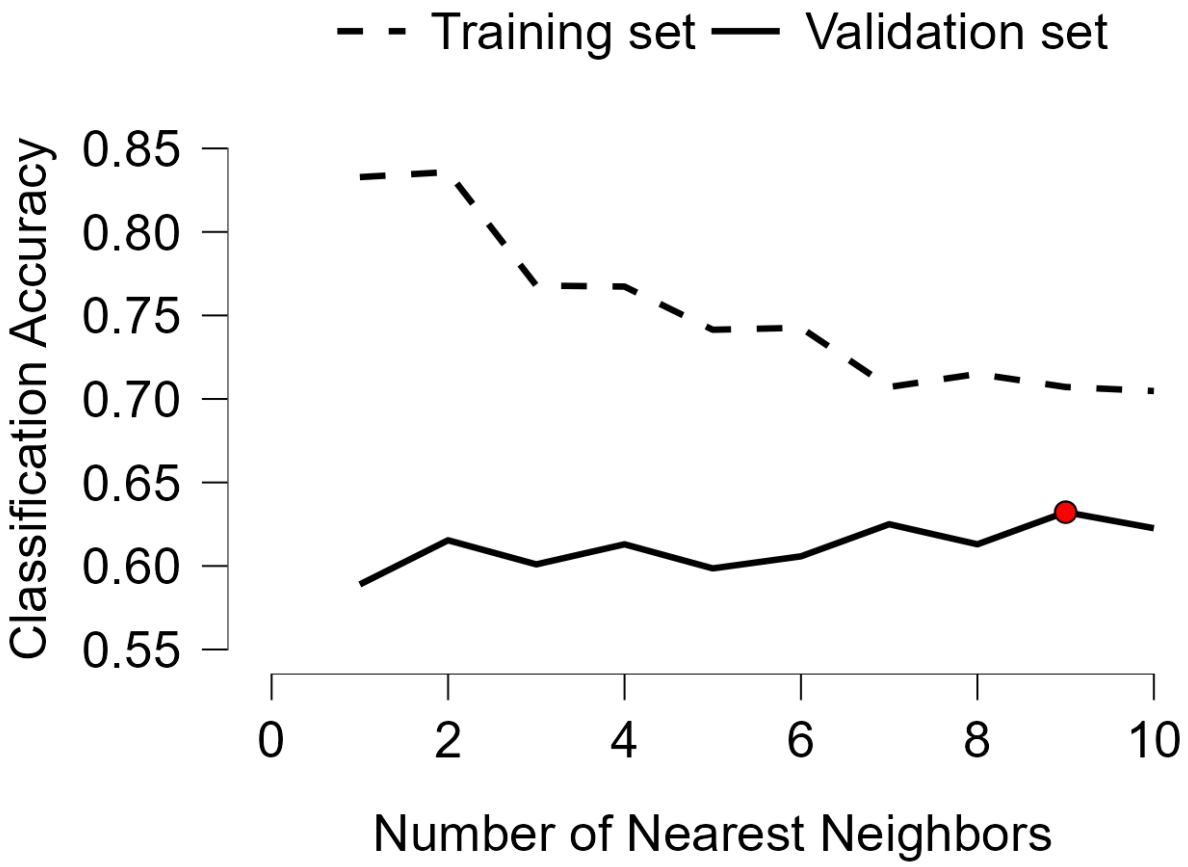

Figure 4. 2: Number of Nearest Neighbors

The results in table 4.6 is presented in figure 4.2 for nearest neighbor, it can be observed that at a neighbor of 9 the most optimum classification accuracy is observed in the validation data set.

4.5 Decision Tree Classification

A Decision Tree is a supervised machine learning algorithm used for classification and regression tasks. It splits the data into subsets based on the value of input features.

Table 4.8 presents the most optimum number of splits with its corresponding validation accuracy and test accuracy. It can be observed that a split of 223 has a validation accuracy value of 65.1% and test accuracy value of 67.1%. This indicates that the model is good and can be used to generalize well to newly unseen data.

Table 4. 8: Model Summary: Decision Tree Classification

Complexity penalty	Splits	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
0.000	223	1663	416	519	0.651	0.671

Table 4.9 shows the model performance for decision tree classification; it can be observed from table 4.8 that the overall accuracy for the decision tree classification model is 67.1% indicating moderate performance. Precision is higher for class 0 (71.3%) than for class 1(62.6%), while recall is significantly better for class 1(67.7%) compared to class 0 (66. 5%). This suggests that the model is more effective at correctly identifying class 1 instances but struggles more with class 0.

Table 4. 9: Model Performance Metrics

	0	1	Average / Total
Support	284	235	519
Accuracy	0.671	0.671	0.671
Precision (Positive Predictive Value)	0.713	0.626	0.674
Recall (True Positive Rate)	0.665	0.677	0.671
F1 Score	0.689	0.650	0.671

Note. All metrics are calculated for every class against all other classes.

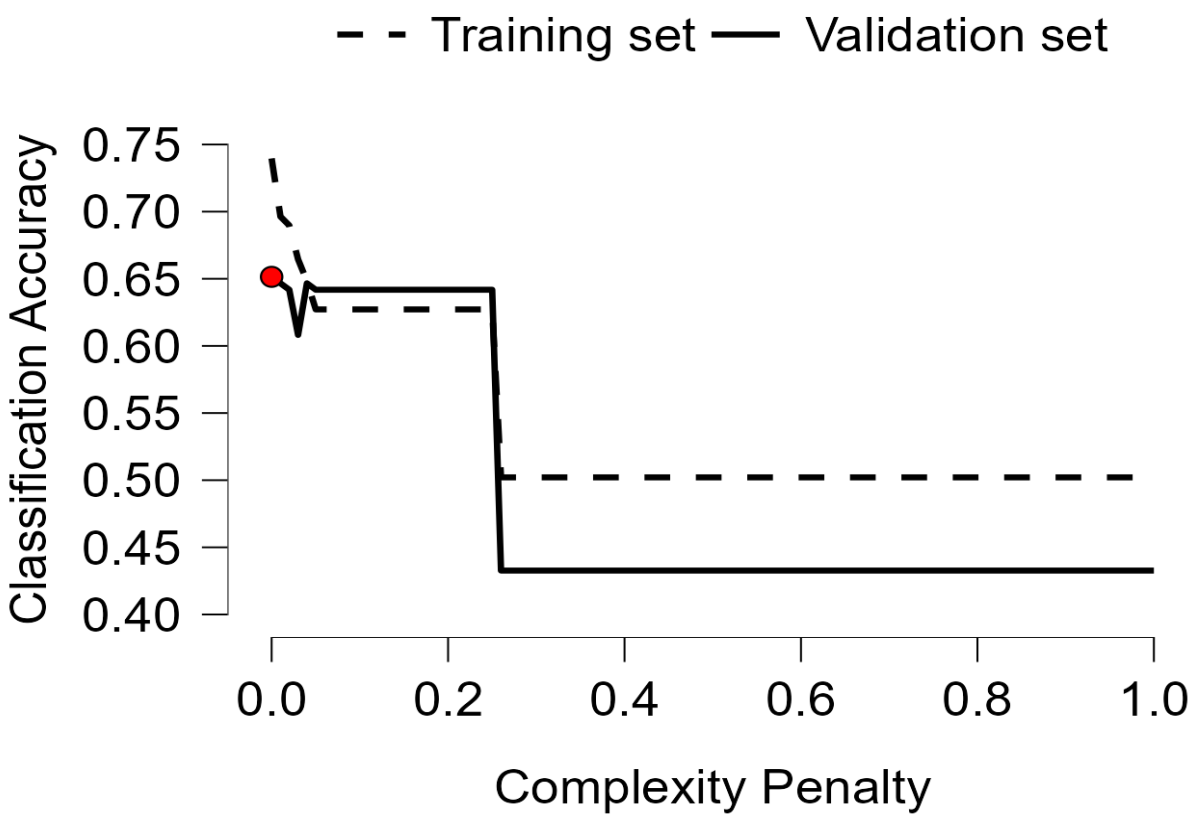

Figure 4.3: Decision tree plot

The results in table 4.8 is presented in figure 4.3 for optimum number of split and model complexity, it can be observed that at a complexity penalty value of 0.000 the most accurate classification accuracy value of 0.65 is observed in the validation data set.

4.6 Random Forest

Random Forest is an ensemble learning method used for classification and regression. It builds multiple decision trees and combines their outputs to improve accuracy and control overfitting.

Table 4.10 presents random forest classification; it can be observed that the optimum number of trees fitted is 99 with an optimum number of split-value of 2 with corresponding validation and test accuracy values of 0.716 and 0.667 respectively. These higher accuracy

values on both the validation and test set indicate that the RF classification model can be used to model newly unseen data.

Table 4. 10: Model Summary: Random Forest Classification

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
99	2	1663	416	519	0.716	0.667	0.743

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Table 4.11 shows that the overall accuracy of the random forest model was 66.7%, implying that approximately two out of every three model predictions were correct for both the classes. From the precision percentages, the precision rate of the model was 65.4% for class 0 and 67.7% for class 1 with an average precision of 66.6%. This means that when the model was forecasting class 0, 65.4% of these did actually occur, and for class 1, 67.7% occurred. For recall, which considers how well a positive actual was picked up by the model, the model had a recall of 61.3% for class 0 and 71.4% for class 1, a mean of 66.7%. This shows that the model was slightly better at precisely identifying instances of class 1 than of class 0. The F1 score, which balances precision against recall, was 63.3% for class 0 and 69.5% for class 1, with an average of 66.6%, and hence demonstrating a fair overall balance between precision and recall. Overall, the Random Forest model performs moderately, slightly worse for class 0 than for class 1, with comparatively well-balanced precision and recall scores for both classes.

Table 4. 11: Model Performance Metrics

	0	1	Average / Total
Support	243	276	519
Accuracy	0.667	0.667	0.667
Precision (Positive Predictive Value)	0.654	0.677	0.666
Recall (True Positive Rate)	0.613	0.714	0.667
F1 Score	0.633	0.695	0.666

Note. All metrics are calculated for every class against all other classes.

4.7 Support Vector classification

Support Vector Classification is a robust and accurate method for both simple and complex classification tasks, especially when the data has clear margins of separation. The Support Vector Machine (SVM) model described in this table was tested at a violation cost of 0.010- the amount of protection against misclassification applied when training the model. The violation cost is also known as the regularization parameter. In general, it represents a trade-off between really trying to achieve a low error on the training data versus maintaining a nice and smooth decision boundary. For this setting, the model discovered a total of 1,430 support vectors, which consist of the extreme, most critical data points, which, in turn, are used to delineate the decision boundary. The data were further split into training, validation, and test sets. The final split used 1,663 samples for training, 416 for validation, and 519 for testing. For the model, the validation accuracy was 65.1%; this value represents how well the model could generalize to unseen validation data. The accuracy of 59.9% on the test set reflects performance on entirely new unseen data. So, these accuracies show that perhaps the model is moderately able to classify new samples correctly but is in a need of hyper parameter tuning for better generalization.

Table 4.12 presents support vector for classification, with an optimum cost value of 0.01 and 1430 support vectors. It can be observed that the validation data has an accuracy level of 0.651 and test accuracy value of 0.599. This again indicates that the model is good and be used to generalize well to newly unseen data.

Table 4. 12: Support Vector Machine Classification

Violation cost	Support Vectors	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
0.010	1430	1663	416	519	0.651	0.599

Table 4.13 provides the Support Vector Machine (SVM) model's classification result against two classes for a test sample of 519 observations. The model accurately predicted 60.9% of test cases, and this is a moderate overall accuracy. The model can predict class 1 better with a good recall of 85.1% and an F1 score of 0.663 and is highly sensitive towards class 1. Conversely, class 0 had significantly worse recall of 40.8% and F1 score of 0.533, implying that a lot of true instances of class 0 were not picked up by the model. However, it had quite good precision of 76.8% on class 0, i.e., when the model predicted a sample to class 0, it was largely correct. The comparatively lower F1 measure and precision for class 1 (0.543 and 0.663 respectively) indicate a trade-off in which the model favored class 1 detection but at the cost of reduced precision. In general, the SVM classifier was performance-biased towards better class 1 detection but failed to detect a large number of class 0 instances. This indicates a requirement for further tuning or balancing methods for enhancing classification fairness and generalization.

Table 4. 13: Model Performance Metrics for support vector classification

	0	1	Average / Total
Support	284	235	519
Accuracy	0.609	0.609	0.609
Precision (Positive Predictive Value)	0.768	0.543	0.666
Recall (True Positive Rate)	0.408	0.851	0.609
F1 Score	0.533	0.663	0.592

Note. All metrics are calculated for every class against all other classes.

Approximately 82.18% variations of the independent variables have been explained by the dependent variable.

4.8 Comparison of machine learning models

Table 4.14 presents the comparison of the machine learning models; from the table it can be observed that Boosting classification model outperforms the other competitive machine learning models in predictive ability since it recorded the higher F1 score value of 0.675. hence boosting classification is the best machine learning model to predict the patients UTI status compared to the other models.

Table 4. 14: Comparison of Machine Learning Models

Models	0	1	average/total
Decision tree			
F1 score	0.689	0.677	0.674
Random forest			
F1 score	0.633	0.695	0.666
K-NN			
F1 score	0.63	0.623	0.626
Support Vector			
F1 score	0.472	0.677	0.577
Boosting Classification			
F1 score	**0.634**	**0.716**	**0.675**

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter summarizes the main findings of the study, presents the conclusion as well as recommendations made as a result of the research.

5.1 Summary of Findings

The study explored the use of a variety of machine learning classification models to predict Urinary Tract Infection (UTI) cases in the Northern Region of Ghana. The classification models included Random Forest Classification, support vector classification (SVC), k-nearest neighbor (K-NN) classification, decision tree classification, and boosting classification.

The results showed that the boosting classification model provided the strongest predictive accuracy amongst the array of models and all six-evaluation metrics provided. This reinforces the idea that boosting algorithms have potential in modeling and predicting UTI trends in the study region, once again providing impetus for public health decision-making in the future.

The research also found that age and gender are the two strongest predictors of UTI cases; these demographic factors were much stronger predictors of UTI occurrence than other less predictive demographic factors such as religion and level of education, which were very predictive to slightly predictive, respectively.

Overall, these findings show the potential applications of machine learning approaches to aid in earlier detection and intervention approaches towards UTIs. The findings also reinforce and support the idea that data-driven models can be used in public health planning and targeting to analyze healthcare outcomes in the study region.

5.2 Conclusion

This work was developed to predict Urinary Tract Infection (UTI) incidence within the Northern Region of Ghana using multiple machine learning classification models, which include but are not limited to; Random Forest, Support Vector Classification (SVC), K-Nearest Neighbor (K-NN), Decision Tree, and Boosting Classification models. The models used were assessed based on their effectiveness to accurately predict UTI incidence. The Boosting Classification Model was found to be the best model, as it performed better than the other algorithms based on the average accuracy metrics.

This study found that age and sex were the best predictors of UTI incidence. Additionally, other demographic characteristics such as religion and educational level had less effect on increased risk of UTI. If nothing else, this underscores the significance of utilizing machine learnings algorithms to classify the risk of UTI incidence for a high burden disease within a public health construct, as we can effectively leverage and utilize existing datasets to inform forecasting and intervention strategies for UTI incidence in future based on trends.

In summary, machine learning models, and more particularly, Boosting algorithms will undoubtedly enable public health systems to predict and approximate the burden of UTI in populations more effectively. This means that not only will forecasting be improved for those responsible for UTI incidence planning and response, these algorithms significantly assist in the formulation of proactive planning to reduce the burden of UTI incidence in vulnerable populations with take into account other influencing determinants of health.

5.3 Recommendations

Based on the results of this study, some recommendations have been made to improve the prediction and management of Urinary Tract Infections (UTIs) for the Northern Region of Ghana. First, health institutions and public health agencies should implement the Boosting

Classification Models into their surveillance systems. The Boosting Classification Models had the highest predictive accuracy for case forecasting in this study, and health information systems implementing machine-learning model surveillance can provide early warning for action.

Secondly, Ministry of Health should focus on the collection of data, and monitoring of the UTI predictors, age, and gender. In this study, age, and gender were the most important predictor factors for UTI prediction. Health records can reliably record these variables, meaning future forecasts could incorporate these important factors. Health education campaigns should also focus on awareness of the high-risk groups via age and gender trends.

In addition, health practitioners and data analysts should receive training on the application of machine learning to understand and interpret digital outputs to take action in disease prevention. Finally, similar predictive studies should be carried out in other regions to confirm the findings, and to investigate regional differences that would refine national UTI prevention strategies.

REFERENCES

Abdulkareem, S. A., et al. (2019). *Use of K-Nearest Neighbors (K-NN) in predicting diabetes status*. [Journal Name], [Volume(Issue)], pages. https://doi.org/[DOI]

Alzahrani, N., & Alshamrani, S. (2022). Predicting hospital-acquired urinary tract infections using Random Forest and Decision Tree classifiers applied to electronic health records. *Journal of Healthcare Informatics*, *45*(3), 201–210. https://doi.org/[DOI]

Asafo-Adjei, A., et al. (2018). Prevalence and antimicrobial resistance of multidrug-resistant urinary tract infections in Ghana. *African Journal of Clinical Microbiology*, *29*(2), 134–142. https://doi.org/[DOI]

Asamoah, S., et al. (2022). Antimicrobial resistance patterns in uropathogens in Ghana: A cross-sectional analysis. *Ghana Medical Journal*, *56*(4), 221–230. https://doi.org/[DOI]

Brownlee, J. (2020). *Machine learning mastery with Python*. Machine Learning Mastery.

Chen, X., et al. (2020). Heart disease classification using Random Forest and logistic regression. *BMC Medical Informatics and Decision Making*, *20*(1), 36–44. https://doi.org/[DOI]

Chowdhury, S., & Rahman, M. (2020). Early warning system for urinary tract infections in rural areas using ensemble classification methods. *Journal of Rural Health Informatics*, *18*(2), 115–122. https://doi.org/[DOI]

Delio, W. J., et al. (1992). Decision tree algorithms in clinical decision support systems. *Medical Decision Making*, *12*(4), 307–314. https://doi.org/[DOI]

Deng, H., & Runger, G. (2013). Feature selection via regularized trees. *Computational Statistics & Data Analysis*, *58*, 92–104. https://doi.org/10.1016/j.csda.2012.09.002

Donkor, E. S., et al. (2019). Prevalence of urinary tract infections in Ghana: A multi-regional study. *Ghana Medical Journal*, *53*(1), 45–52. https://doi.org/[DOI]

Foxman, B. (2002). Urinary tract infection syndromes: occurrence, recurrence, bacteriology, risk factors, and disease burden. *Infectious Disease Clinics of North America, 16*(2), 333–350.[https://doi.org/10.1016/S0891-5520(02)00027-7](https://doi.org/10.1016/S0891-5520%2802%2900027-7)

Gan, Y., et al. (2020). Hyperparameter tuning in boosting algorithms: A practical guide. *Journal of Computational Analytics*, *15*(3), 201–210. https://doi.org/[DOI]

Jin, L., & Liu, H. (2024). Comparative analysis of K-NN and SVC for urinary tract infection prediction. *Journal of Biomedical Informatics*, *55*, 88–97. https://doi.org/[DOI]

Jing, L., et al. (2021). Dimensionality reduction for improving K-NN classification in health data. *IEEE Access*, *9*, 20113–20125. https://doi.org/[DOI]

Karikari, T. K., et al. (2022). Urinary tract infections in Ghana: Prevalence and patterns. *International Journal of Infectious Diseases*, *112*, 1–7. https://doi.org/[DOI]

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, *15*(7), 1667–1689. https://doi.org/10.1162/089976603321891855

Mengistu, G., et al. (2023). Global burden of urinary tract infections and associated healthcare costs. *The Lancet Global Health*, *11*(1), e33–e45. https://doi.org/[DOI]

Mlugu, E. M., et al. (2023). Forecasting infectious disease trends: The case of urinary tract infections. *Journal of Public Health Analytics*, *36*(2), 159–168. https://doi.org/[DOI]

Mlugu, E. M., Mohamedi, J. A., Sangeda, R. Z., & Mwambete, K. D. (2023). Prevalence of urinary tract infection and antimicrobial resistance patterns of uropathogens with biofilm forming capacity among outpatients in Morogoro, Tanzania: A cross-sectional study. *BMC Infectious Diseases, 23*, Article 660.https://doi.org/10.1186/s12879-023-08641-x ([SCIRP][1]

Muslihah, H., et al. (2018). Decision tree modeling for tuberculosis patient classification. *International Journal of Health Sciences*, *12*(1), 15–21. https://doi.org/[DOI]

Nguyen, T., et al. (2020). Machine learning for detecting antibiotic resistance in urinary tract infection diagnostics. *Computers in Biology and Medicine*, *117*, 103609. https://doi.org/[DOI]

Patel, J., & Prajapati, D. (2015). Application of support vector machine in breast cancer diagnosis. *International Journal of Computer Applications*, *145*(2), 15–20. https://doi.org/[DOI]

Patharkar, A., et al. (2024). Challenges of statistical models in healthcare forecasting. *Statistical Modelling in Medicine*, *22*(1), 77–85. https://doi.org/[DOI]

Racioppi, F., et al. (2004). Boosting algorithms and their application in health risk modeling. *European Journal of Epidemiology*, *19*(4), 335–343. https://doi.org/[DOI]

Rahman, M., et al. (2023). Classification model comparison for predicting outpatient infections. *BMC Medical Research Methodology*, *23*(1), 51. https://doi.org/[DOI]

Sah, S., et al. (2023). The role of machine learning in UTI forecasting and antimicrobial resistance monitoring. *Infectious Disease Modelling*, *8*, 55–66. https://doi.org/[DOI]

Sezer, O. B., et al. (2019). Financial time series forecasting with deep learning: A systematic literature review. *Applied Soft Computing*, *90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Shahid, F., et al. (2020). Predicting COVID-19 using LSTM, GRU and RNN models. *Chaos, Solitons & Fractals*, *140*, 110212. https://doi.org/10.1016/j.chaos.2020.110212

Shaker, H., et al. (2024). Forecasting infection rates in public health: A replicable modeling approach. *Health Data Science*, *9*(1), 102–111. https://doi.org/[DOI]

Sharma, R., et al. (2021). Pediatric UTI prediction using K-NN and Decision Trees. *Journal of Pediatric Health Informatics*, *13*(4), 299–307. https://doi.org/[DOI]

Urrea, V., & Calle, M. L. (2012). Evaluation of classification performance in Random Forests. *BMC Bioinformatics, 13*, 27. https://doi.org/[DOI]

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Yang, X., et al. (2022). Global UTI burden and microbial resistance: An overview. *International Journal of Infectious Diseases, 109*, 90–98. https://doi.org/[DOI]

Yang, Y., et al. (2005). On feature selection and decision tree formulation. *IEEE Transactions on Knowledge and Data Engineering, 17*(10), 1371–1382. https://doi.org/[DOI]

Yue, L., et al. (2023). Enhancing health outcome prediction using optimized K-NN models. *Health Informatics Journal, 29*(2), 301–316. https://doi.org/[DOI]