HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

INVESTIGATION OF MACHINE LEARNING MODELS FOR FOODBORNE DISEASE CLASSIFICATION

By:

WULETAWU IYASU FARACHO

HAWASSA, ETHIOPIA

JUNE, 2024

THESIS ON

INVESTIGATION OF MACHINE LEARNING MODELS FOR FOODBORNE DISEASE CLASSIFICATION

BY

WULETAWU IYASU

Advisor: Dr.DEGIF TEKA (PhD)

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE

HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

HAWASSA, ETHIOPIA

JUNE, 2024

HAWASSA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

EXAMINER'S APPROVAL SHEET-I

This is to certify that the thesis entitled "**Investigation of Machine Learning Models for Foodborne Disease Classification**" submitted in partial fulfillment of the requirements for the degree of Master's with specialization in Computer Science, the Graduate Program of the Department/School of Informatics, and has been carried out by Wuletawu Iyasu. Therefore we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

Dr. Degif Teka

_____	_____	_____
Name of major advisor	Signature	Date

_____	_____	_____
Name of co-advisor	Signature	Date

HAWASSA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

EXAMINER'S APPROVAL SHEET-II

We, the undersigned, members of the Board of Examiners of the final open defense by Wuletawu Iyasu have read and evaluated his/her thesis entitled "**Investigation of Machine Learning Models for Foodborne Disease Classification** ", and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

_____	_____	_____
Name of Major Advisor	Signature	Date
_____ _	_____	_____
Name of Internal Examiner-I	Signature	Date
_____	_____	_____
Name of Internal Examiner-II	Signature	Date
_____	_____	_____
Name of External examiner	Signature	Date
_____	_____	_____
SGS Approval	Signature	Date

ACKNOWLEDGMENT

First and foremost, I would like to express my profound gratitude to Almighty God for being with me in every situation and for keeping me well so that I could complete my research. Next, I would like to express my deepest gratitude and appreciation to my principal advisor, Dr. Degif Teka (PhD) for his excellent and constructive comments, suggestion and remarkable follow up in every element of this research and I will never forget his sincerity and support. He is the ideal role model and the exact definition of a leader.

I am grateful for my parents whose constant love and support, which keep me inspired and self-assured. They believed in me, and as a result, I achieved success. My next thanks go out to my Instructors at Hawassa University Faculty of informatics for helping me in my academic endeavors in a variety of ways, as well as to my classmates for contributing their suggestions about potential research topics and supporting me during all of my experiences. And also I would like to thank Wolaita Sodo University who sponsored me to study Msc degree at Hawassa University and my colleagues at Wolaita Sodo University for their good wishes in my success.

Finally, I'd like to express my gratitude to the doctors, nurses, and other staff members at the Gesuba Primary Hospital and Wolaita Sodo University Comprehensive Specialized Hospital for their consultation and knowledge sharing on foodborne diseases and gave permission to collect secondary data.

STATEMENT OF THE AUTHOUR

I hereby declare that this MSc thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis / have been duly acknowledged.

Name: WULETAWU IYASU FARACHO

Signature: _____

Place: Institute of Technology, Hawassa University, Hawassa
Date of Submission: _____

ABSTRACT

Foodborne disease is a disease that has a high prevalence in low and middle income countries around the world. There are many people affected by foodborne disease in Ethiopia, due to various causes. There are high burdens of infection; the control of most foodborne diseases in Ethiopia is in its infancy due to a lack of technology that can classify foodborne diseases easily in order to support healthcare professionals for better diagnoses. There is a lack of study conducted to classify the foodborne diseases which are common in Ethiopia. It is in view of this facts, the study aims to undertake an investigation on the topic and fill the research gap observed using machine learning model which can learn from past data, identify patterns and make decisions with a minimal human intervention. These applications in the healthcare and biomedical domain are popular for the early detection of diseases and help to make a better diagnosis. This study focuses on foodborne diseases, some of the prevalent foodborne illnesses in Ethiopia, selected in consultation with medical experts. To achieve the objective of the study the researcher used an experimental research design and mixed research approach (both quantitative and qualitative). For this study, secondary data of foodborne diseases were collected from Hospitals, and to perform most of the research activities such as data pre-processing, analysis, model training, and testing, python programming is used, and to design a conceptual model, Edraw max is used based on its good features. After preprocessing the collected data, the researcher trained a model using state-of- art machine learning algorithms like Decision Tree, Random Forest, XGBoost and Stacking ensemble learning method. Based on the experiment conducted, the Stacking ensemble learning method model outperforms the others with an accuracy of 98.1%, followed by Random Forest, XGBoost, and Decision Tree with accuracy of 97.5%, 96.9%, and 96.5% respectively. The result obtained by the study depicts that, the Stacking ensemble learning model is suitable for diseases classification.

Keywords: *Disease Classification, Disease Prevalence, Machine Learning Model, Stacking ensemble learning method, foodborne Disease.*

Table of Contents

List of Acronyms

AI	Artificial Intelligent
ANN	Artificial Neural network
AUC	Area under Curve
DALYs	Disability-Adjusted Life Year
DNN	Deep Neural Network
DT	Decision Tree
FBD	Food Borne Disease
GPH	Gesuba Primary Hospital
HIPAA	Health Insurance Portability and Accountability Act
IDE	Integrated Development Environment
IRB	Institutional Research Review Board
KNN	K-Nearest Neighbors
ML	Machine Learning
MOH	Ministry of Health
MRN	Medical Record Number
NB	Naïve Bayes
NN	Neural Network
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector machine
WHO	World Health Organization
WSUCSH	Wolaita sodo university comprehensive specialized hospital
XGBOOST	eXtreme Gradient Boosting

List Table

List of Figures

List of Equation

Chapter One

Introduction

1.1 Background of the study

Access to safe and nutritious food is a fundamental human right and essential for well-being[1].However, foodborne diseases, caused by consuming contaminated food or water, pose a significant threat to public health globally[2] .They are usually associated with contaminated foods and pathogens contained in foods. According to a World Health Organization (WHO) report, 600 million people worldwide suffered from diseases caused by eating contaminated food every year, of whom 420000 die[3]. In Ethiopia, food safety is a major concern due to lack of infrastructure and basic pre-requisites for food safety such as clean water and environment, washing facilities, compounded by limited implementation of food safety regulations, and a lack of incentives for producers to improve food safety [4]. Frequent foodborne illness outbreaks both domestically and internationally pose a major threat to social stability and public health, making them a global public health and food safety concern.

Machine Learning (ML) is a subfield of artificial intelligence that can learn from experience, identify patterns and make decisions with minimal human intervention using its algorithms. Classification is one of the machine learning tasks that is used to classify different patterns based on past data. These applications in the healthcare domain have led to early detection of disease and helped to make a better diagnosis [5].

Several studies are conducted on disease classification using machine learning based on symptoms and image analysis. Among them, the classification of heart disease, kidney disease, stroke, breast cancer, and pneumonia disease are most commonly investigated in the literature using machine learning algorithms [6]. But there is lack of research conducted around the world on foodborne disease classification in order to detect these diseases early before spreading and identify the specific disease from other diseases which shares symptoms in common. From the reviewed literature there is an attempt to classify foodborne diseases such as Salmonella (which can manifest as Typhoid fever), Norovirus (known to cause acute gastroenteritis with vomiting and diarrhea), E. coli (causing a range of symptoms from mild to severe cramps and diarrhea, sometimes with

blood), and Vibrio parahaemolyticus (a bacterial infection acquired through consuming raw or undercooked seafood). However, the most common foodborne diseases in Ethiopia are not included in the previous study [6]. Thus, there is a need for a classifier model that can classify the most common foodborne disease that occur in Ethiopia in order to analyze the prevalence of a disease in a specific area and ease the detection of specific diseases early to make a better decision while diagnosing. Therefore, the researcher is motivated to investigate a machine learning models to classify foodborne diseases which are common in Ethiopia so as to fill the identified gaps in the previous studies.

To make a classification in machine learning, it needs datasets to learn from the data and then classify the expected outcomes from unseen data. So, we collected the data on the most common foodborne diseases from the Hospitals. Then, the collected datasets were preprocessed in such a way that it is appropriate for the selected machine learning algorithms. Then, the models were trained and their performance is evaluated by unseen data. Finally, a suitable algorithm is selected to develop a model based on the experiment conducted. To preprocess the dataset and develop the model, we used python programming language.

The proposed study mainly focused foodborne diseases with high prevalence in the study area, particularly Typhoid fever, Giardiasis, and Amoebiasis, which are also common throughout Ethiopia. It is essential to recognize that these prevalence rates can fluctuate across different regions within Ethiopia. The varied cultural practices across Ethiopia likely contribute to differing prevalence of foodborne diseases in various regions[7]. The proposed study significantly helps medical doctors to make a proper diagnosis of specific foodborne diseases at less time.

There are many machine learning algorithms such as Logistic Regression (LR), Naïve Bayes (NB), k-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machines (SVM), Random Forest (RF)which are discussed in the literature[6]. For the proposed model development, the researcher used the state of art machine learning algorithms such as Decision Tree, Random Forest and eXtreme Gradient Boosting (XGBoost) because those algorithms are promising algorithms so as to have a model with a good performance. In addition, the algorithm like Random Forest is used to handle data imbalance[8]. Additionally, the researcher has implemented a stacking ensemble method. This approach leverages the strengths of individual models to potentially achieve superior classification performance[9]. According to the results of the experiment, the stacking ensemble learning method model performs better than the others with an accuracy of 98.1%, followed by

RF, XGBoost and DT with accuracy of 97.5%, 96.9%, and 96.5% respectively. The study's findings show that the stacking ensemble learning model is effective for foodborne diseases classification.

1.2 Statement of the Problem

Foodborne diseases pose a significant threat to public health, with millions of people worldwide suffering from illnesses caused by consuming contaminated food[2]. According to a recent World Bank analysis, the cost of treating these illnesses added an additional $3.5 billion to the productivity losses in Africa alone in 2016 as a result of contaminated food[4]. In Ethiopia, the occurrence of foodborne diseases is a major concern due to inadequate infrastructure, lack of basic prerequisites for food safety, and limited implementation of food safety regulations[7].

According to the World Health Organization (WHO), foodborne diseases can cause a wide range of symptoms, from mild gastroenteritis to life-threatening conditions such as meningitis and cancer. These diseases can also cause long-term health problems, including disability, delays in physical growth, mental stress, blindness, and cognitive growth. The WHO also notes that foodborne diseases can have a significant impact on a person's ability to go to school and work, thus perpetuating the cycle of disease and poverty[10].Furthermore, a report by the Global Alliance for Improved Nutrition (GAIN) states that foodborne diseases can also be a cause of other chronic diseases like heart disease and kidney disease. However, despite these high burdens of diseases, the control of most foodborne diseases in Ethiopia is in its infancy. One of the causes might be associated to the lack of technology that can easily classify foodborne diseases to support health care professionals in order to identify the foodborne disease which shares symptoms in common and to make a proper diagnosis of disease early. It is important to take countermeasure before spreading and getting the situation worse. Moreover, In Ethiopia, a significant portion of the population lacks access to healthcare facilities and diagnostic testing. This, coupled with the widespread availability of antibiotics without prescriptions, creates a critical issue of potential overuse and misuse of these essential medications. Studies[11], [12] have shown that the extent of dispensing antibiotics without a prescription in Ethiopia is high as 67.7% are dispensed without a doctor's visit , often for self-diagnosed illnesses. Additionally, the majority of foodborne illnesses in Ethiopia are treated clinically without laboratory confirmation, mainly in health centers[13]. This is a significant issue as it can lead to the incorrect diagnosis of foodborne illnesses and the

inappropriate use of medication. The previous study attempts to classify foodborne diseases namely Salmonella, Norovirus, Escherichia coli, and Vibrio parahaemolyticus which are common in china[14]. But there is a lack of adequate studies conducted to classify the foodborne disease which are common in Ethiopia as the diseases are different and similarly the symptoms. Thus, there is a need for a technology that helps to achieve the plan of the federal MoH(ministry of health) which are improving the quality of health services, increasing access to health care, and strengthening the health system in the country[15] by easily classifying foodborne diseases which are common in Ethiopia. Therefore, to solve the aforementioned problem and fill the research gap the researcher is motivated to investigate a machine learning model that can classify the foodborne diseases which are common in Ethiopia.

1.3 Research Questions

RQ1. What are the important attributes that are used to classify foodborne diseases?

RQ2. Which machine learning model best suit for classifying foodborne diseases?

1.4 Objectives of the study

1.4.1 General Objective

The prime objective of this study is to develop a classifier model for foodborne diseases using machine learning.

1.4.2 Specific Objective

To achieve the general objective, this research targets the following specific objectives:

- To review related literature for better understanding of the problem and research design
- To collect and preprocess the data
- To identify important attributes that are used to classify foodborne diseases.
- To select a suitable model for foodborne disease classification
- To train and develop a model for foodborne disease classification.
- To evaluate and compare the performance of the models.

1.5 Significance of the Study

The main significance of the proposed study is to support healthcare professionals so as to make a better diagnosis and early detection of the diseases by easily identifying foodborne diseases using a machine learning model. Also, it is helpful for policy makers to take proactive measures before the spreading of the disease, plan for the future, and to make better decisions regarding foodborne diseases. The proposed study also helps the researchers and academicians to make further studies, analyses, and investigations on foodborne diseases.

1.6 Beneficiaries of the Proposed Study

The proposed study primarily helps in the quick and simple diagnosis-making process for physicians. The ability to diagnose quickly and accurately can provide good service to patients and to safe-guard their lives. Moreover, the proposed study benefits the hospitals and the country at large to defeat the foodborne diseases early on and prevent disease outbreaks..

The beneficiaries of the proposed study are:

- Doctors and Nurses: to easily diagnose foodborne diseases with the help of a machine learning model
- MOH and Policymakers: to plan for the future, regarding foodborne diseases and to take proactive measures before the spreading of diseases.
- Hospitals and the researchers: for easy analysis of data regarding foodborne diseases and to facilitate daily routine workflow.
- Patients: to safeguard their life by quickly making a proper diagnosis.

1.7 Scope and Limitation of the Study

The proposed study mainly focused on foodborne diseases which are most common in Ethiopia. Among those diseases, the researcher selected the diseases which are infectious and has a high prevalence in study area such as Typhoid fever, Amoebiasis and Giardiasis. Before collecting the confirmed cases of foodborne infectious from the hospitals the researcher had made a consultation with medical experts so as to have a good understanding of the problem domain and select the Hospitals where the essential data for proposed study are to be collected. Based on consultations made with medical experts and the researchers, the data were collected from Wolaita sodo university comprehensive specialized hospital and Gesuba primary hospital. The collected sample

data were taken from the year 2012 to 2015 E.C and covers all age group and gender in order to address and analysis the prevalence of the diseases in all populations. The proposed study does not fully address all of the foodborne diseases which are prevalent in Ethiopia due to budget and availability of the data constrain.

1.8 Ethical Consideration

Ethics of research here refers to the morals of investigation or intervention about minimal abuse or disregard of social and psychological wellbeing of persons, community, and animals participating in the research work. The researcher obtained ethical clearance from Hawassa University College of Medicine and Health sciences Institutional Review Board (IRB) to collect the secondary data from the Hospitals as shown in Appendix G. The issue of confidentiality is considered while collecting and analyzing the data to keep the social and psychological well-being of persons by taking the information only for the intended purpose without using any personal identifiers like a patient's name and Medical Record Number (MRN). The Patient's data were protected as per the rules of the Health Insurance Portability and Accountability Act (HIPAA).

1.9 Organization of the Thesis

There are six chapters in this thesis document. The first chapter deals with the background of the study, the statement of the problem, the study's general and specific objectives, the study's scope, and the significance of the study. In Chapter 2 conceptual literature reviews and related work are discussed. In this chapter, the researcher covers the fundamental ideas of machine learning concerning the proposed study as a basic concept and the prevalence of foodborne diseases in Ethiopia. Also, a review of related work that is relevant to the proposed study was investigated. The study methodology such as research approach, research methods, data analysis, and data preprocessing components such as data transformation, finding missing value, and feature selection are all covered in Chapter 3. In chapter 4, proposed model design and development were discussed including the performance evaluation metrics of the models. Chapter five presents the experiment and result discussion of the selected models. Chapter six focused on the conclusion, recommendation, and contribution of the study. Based on the results, the researcher presents the findings and concludes the study by recommending future work.

Chapter Two

Review of Related Literature

2.1 Conceptual Literature Review

In this research different books, journal articles, and conference proceedings have been reviewed based on the relevancy of the research domain, the quality of the journal, and the years of the publication to gain better understanding of the problem and research design.

2.2 Overview of foodborne diseases

Foodborne diseases are illnesses that result from eating food contaminated with harmful agents such as bacteria, viruses, parasites, or chemicals. They can cause a variety of symptoms, ranging from mild to severe, and affect different parts of the body[10]. Typhoid fever, Giardiasis, and Aamebiasis are some of the common foodborne diseases.

- Giardiasis is a parasitic infection caused by the Giardia lamblia parasite. Symptoms include diarrhea ,fatigue , abdominal discomfort, weight loss, nausea, [16].
- Amoebiasis is a parasitic infection caused by the Entamoeba histolytica parasite. Symptoms include diarrhea, vomiting, abdominal cramp, dysentery, weakness [17].
- Typhoid fever is a bacterial infection caused by the Salmonella typhi bacteria. Symptoms include fever, chills, headache ,diarrhea, abdominal flank, muscle or joint pain, cough, loss of appetite.[18].

2.2.1 Worldwide Prevalence of foodborne diseases

According to the World Health Organization (WHO), unsafe food causes 600 million cases of foodborne diseases and 420000 deaths worldwide each year[19]. The global burden of foodborne disease was estimated to be 33 million Disability-Adjusted Life Year(DALYs) in 2010, with 40% of the foodborne disease burden among children under five years of age[20]. WHO has a critical role in taking global leadership in investment and coordinated action across multiple sectors to build strong and resilient national food safety systems and provide consumers with tools to make safe food choices[19].

2.2.2 Prevalence of foodborne diseases in Ethiopia

A systematic literature review conducted by Gazu *et al* [4], reported the occurrence of foodborne diseases (FBD) in Ethiopia between 1990 and 2019. The study identified multiple food safety challenges in Ethiopia, including a lack of infrastructure and basic pre-requisites for food safety, such as clean water and environment, washing facilities, limited implementation of food safety regulations, and a lack of incentives for producers to improve food safety. The study found that FBD affects millions of people each year in Ethiopia, with Typhoid fever(Salmonella) and pathogenic Escherichia coli contributing to 5.7% and 11.6% of the overall pooled prevalence estimates of FBD in Ethiopia, respectively[21]. However, the study also noted that the data on FBD occurrence in Ethiopia is limited and fragmented, highlighting the need for more comprehensive surveillance systems and research to inform policy and future interventions.

Moving beyond FBD, the prevalence of typhoid fever in Ethiopia exhibits variability. A study conducted in Central Ethiopia found a prevalence of 5.0% among febrile patients with symptoms clinically resembling typhoid fever[22]. Furthermore, a recent surveillance report revealed that 29,114 cases of typhoid fever were reported nationally during a specific week[23].

Another infectious disease of concern is amoebiasis. In a countrywide survey, the overall prevalence of Entamoeba histolytica infections (measured by the rate of cyst-passers) was 15.0% among schoolchildren and 3.5% in non-school communities in Ethiopia.[24].

It is essential to recognize that these prevalence rates can fluctuate over time and across different regions within Ethiopia. Public health efforts play a pivotal role in preventing and managing these infectious diseases. As we seek further information, consulting with local health authorities and healthcare providers remains crucial.

2.2.3 Treatment and Management of foodborne diseases

The treatment and management of foodborne diseases depend on the type of pathogen, severity of symptoms, and the patient's health status. The diagnosis and management of foodborne illnesses are based on the history and physical examination. Common symptoms of foodborne illnesses include vomiting, diarrhea (with or without blood), fever, abdominal cramping, headache, dehydration, myalgia, and arthralgia’s [25]. Empiric treatment should focus on symptom management, rehydration if the patient is clinically dehydrated, and antibiotic therapy. Foodborne illnesses should be reported to local and state health agencies; reporting requirements vary among

states[25].

2.3 Overview of Machine Learning

Machine learning is the subfield of Artificial Intelligence (AI) which can learn from experience and make a prediction, classification, and clustering with the minimal human intervention [26]. Now day machine learning has a lot of applications in different fields such as agriculture, healthcare, finance, military, air aviation, transportation, manufacturing etc. To classify the given sets of input variables manually is time-consuming and error-prone, so it needs technology like machine learning to classify, cluster and predict those data within a short period of time. In order to make classification and clustering using machine learning, one has to train a model. Thus, Good quality (preprocessed) data is needed to train a model so as to have output with high accuracy. Different algorithms are used to build ML models depending on the type of data at hand and the domain of the problem. After training the model, the performance of the model is evaluated using different ML evaluation metrics. Finally, an outperformed algorithm is selected to build the model based on experimental results. ML has a lot of applications in healthcare in order to facilitate daily routine work which needs a long time and effort from human beings to perform different tasks like patient disease diagnosis [5].

2.3.1 Types of Machine Learning

Machine learning algorithms can be categorized into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [14] as shown in Figure 2.1.

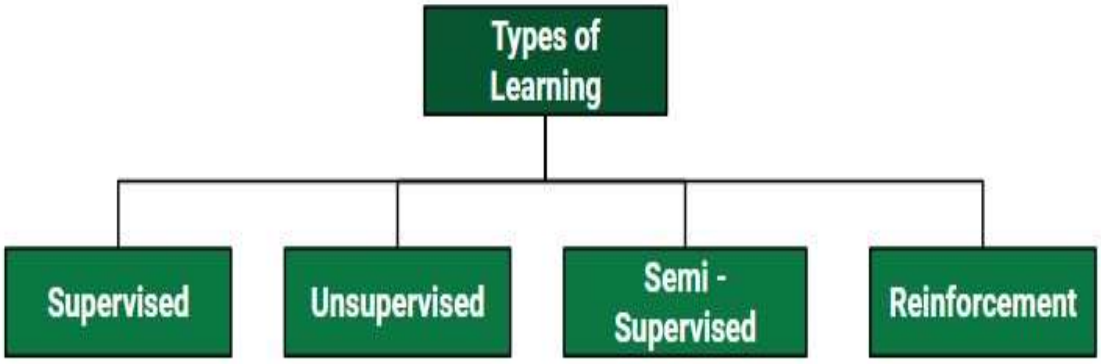


Figure 2.1 Types of machine learning [27]

2.3.1.1 Supervised Learning

Supervised learning refers to a method in which the input and output of the data are already known or labeled to perform classification and prediction task[28]. Unlike unsupervised learning, which deals with unlabeled data, supervised learning algorithms receive training data consisting of inputs (features) and their corresponding desired outputs (labels). By analyzing this data, the model establishes a relationship between the features and labels, enabling it to perform tasks like:

- Classification: Classifying new, unseen data points into predefined categories. Imagine training a model to identify handwritten digits (0-9) based on image data. Presented with a new handwritten symbol, the model would classify it as a specific digit based on the learned patterns.
- Regression: Predicting continuous values based on the input features. For example, a model trained on historical house price data with features like size, location, and market trends could predict the selling price of a new house based on its characteristics.

2.3.1.1.1 Types of Supervised Learning Problems

Supervised learning problems can be broadly categorized into two main types:

Classification Problems: The goal is to predict the class label (category) to which a new data point belongs[29]. Classification problems involve discrete outputs, meaning they fall into distinct categories. Here are some common examples:

- Image Classification: Classifying images into predefined categories like "cat," "dog," "car," etc.
- Spam Detection: Classifying emails as spam or not-spam.
- Disease Diagnosis: Classifying medical data to predict the presence or absence of a specific disease.
- Sentiment Analysis: Classifying text data as positive, negative, or neutral sentiment.

Regression Problems: The aim is to predict a continuous output value for a new data point. Regression problems deal with numerical outputs, and the model learns to map features to continuous values[29]. Here are some common examples:

- Price Prediction: Predicting the future price of a stock, house, or other commodity based on historical data.
- Weather Forecasting: Predicting continuous weather variables like temperature, precipitation, or wind speed.
- Sales Forecasting: Predicting future sales figures based on past sales data and market trends.
- Demand Forecasting: Predicting the demand for a product or service based on various factors.

2.3.1.2 Unsupervised Learning

It is a type of machine learning approach used where the data is not labeled i.e., only the input variable is at hand but the output variable is unknown [30]. We can classify the unsupervised learning problems into two categories: Clustering and Association. Clustering is a type of unsupervised learning in which the data with the same features is grouped into the same category i.e., means bundling the input variables with the same characteristics together while the association is discovering the rules that govern meaningful associations among the data set. In addition to this, unsupervised learning is useful for dimensionality reduction when the

features of the dataset are high.

2.3.1.3 Semi-supervised Learning

It is the combination of both supervised and unsupervised learning. It uses small sorts of labeled data and a large amount of unlabeled data for training and makes classification and clustering based on the given data [31].

2.2.1.4 Reinforcement Learning

Reinforcement learning is also a type of machine learning and it consists of algorithms that are able to perceive and interpret their environment, take actions and learn through trial and error [32]. When the agent performs correctly it receives a reward and penalties if the agent performs incorrectly. . An example is a driverless car.

In this thesis, the researcher has selected supervised machine learning and classification technique for the proposed study in order to classify the foodborne diseases which are most common in Ethiopia so as to help health care professionals to make a better diagnosis.

2.3.2 Machine Learning Algorithms for Classification

Popular and state of art machine learning algorithms such as Decision Tree, Random Forest, and XGBoost were used for classification problems by previous researchers[33]. Those algorithms are promising algorithms to make a classification with high accuracy. Therefore, the researcher focused on those algorithms so as to have a good performance of a proposed model. Additionally, researcher implemented a stacking ensemble method. This approach leverages the strengths of individual models to potentially achieve superior classification performance.

2.3.2.1 Decision Tree

A decision tree is a powerful tool for making decisions based on a tree-like model. It helps us navigate through choices and their potential outcomes. In the context of machine learning, decision trees can be used for both regression and classification problems[34].

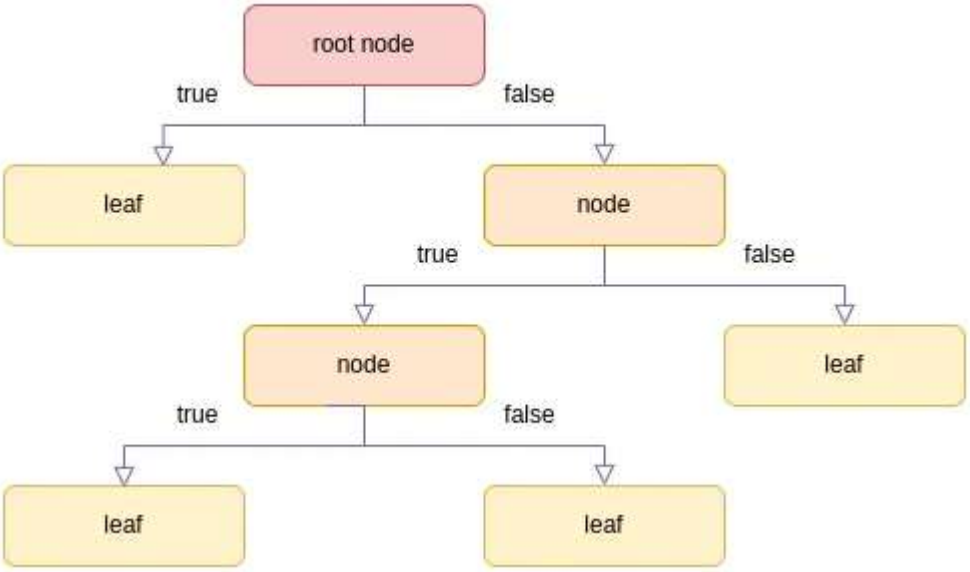


Figure2.2 Shows how the Decision Tree algorithm works[34]

2.3.2.2 Random Forest

The random forest algorithm is one of the popular machine learning algorithms that is used to improve the accuracy of the model by taking all the decisions made by the decision node in the decision tree and then taking the average of those decisions and finally predicting the target output as shown in below Figure 2.3 [35].

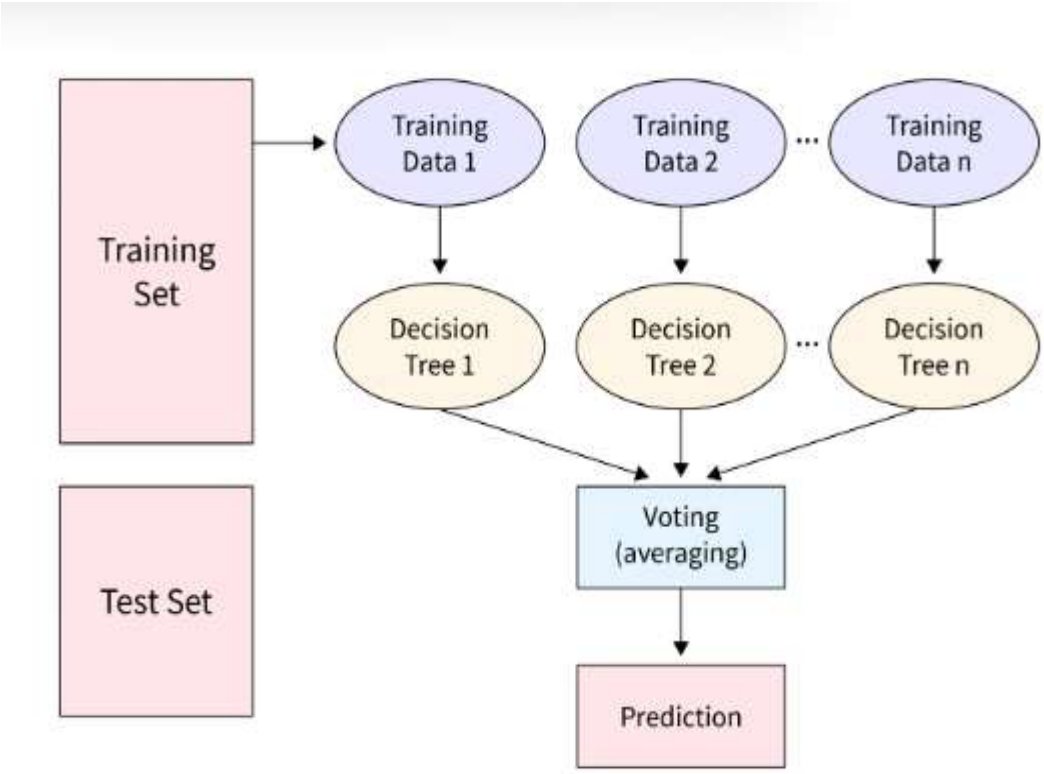


Figure 2.3 Shows how the Random Forest algorithm works[36]

2.3.2.3 eXtreme Gradient Boosting

XGBoost is a gradient boosting approach that uses a decision-tree-based ensemble machine learning algorithm. It uses parallel tree boosting to tackle a wide range of data science issues quickly and accurately. XGBoost is an ensemble tree approach that uses the gradient descent architecture to boost weak learners from the decision tree so as to minimize the loss function of the model [37]. Figure 2.4 depicts how the XGBoost algorithm works.

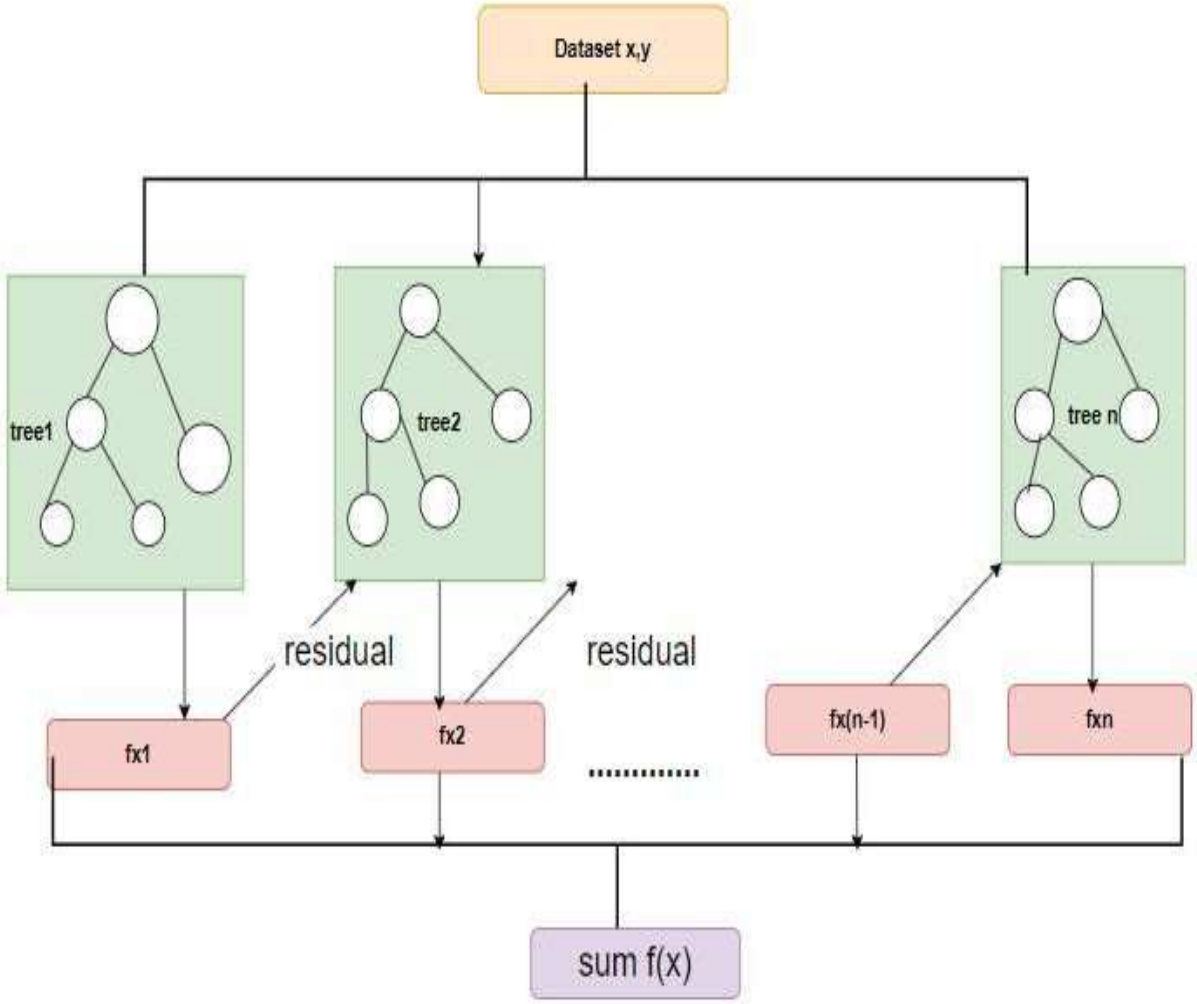


Figure 2.4 Shows how the XGBoost algorithm works [38]

2.3.2.5 Stacking Ensemble Method

Stacking Ensemble models represent a sophisticated machine learning technique that combines multiple predictive models to achieve higher accuracy than any individual model could on its own. This method, also known as stacked generalization, involves training a meta-learner or meta-model to integrate the predictions of several base learners, which are typically diverse in nature. The base learners, also called level-0 models, are trained on the full dataset and make individual predictions. These predictions are then used as input features for the meta-learner, which is trained to make the final prediction. The strength of stacking lies in its ability to capture the unique strengths of each base model, while the meta-learner corrects for their individual weaknesses, leading to a robust

and well-generalized model. The process of stacking not only enhances predictive performance but also provides a framework for understanding how different models can complement each other in a predictive task[9].

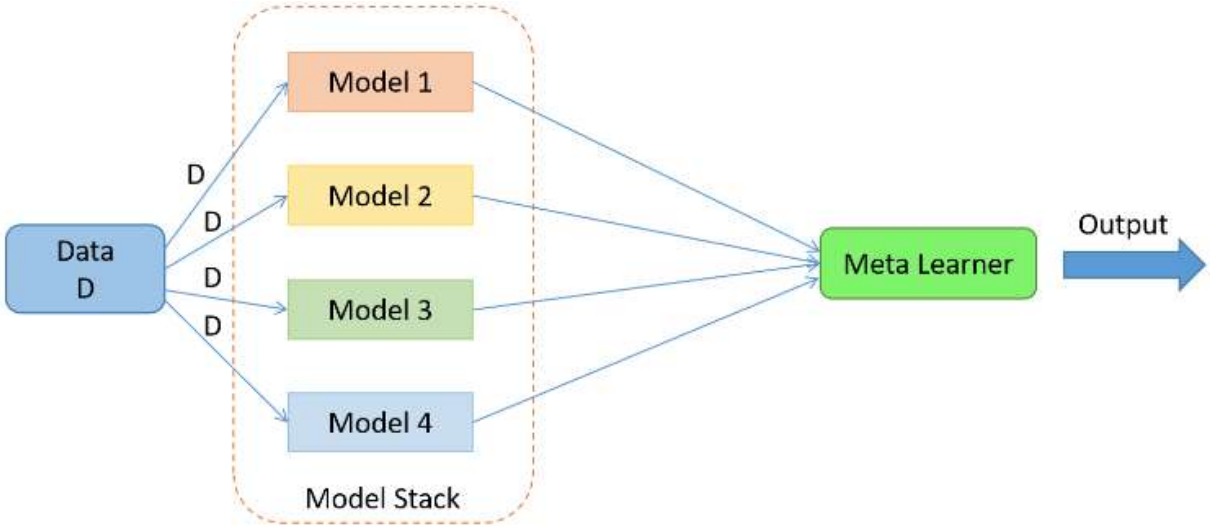


Figure 2.5 Stacking Algorithm[39]

2.4 Review of Related Research Works/Literature

This section presents review of related literature that helps get understanding of the problem from the wider literature and identify the gap. It helps also learn how to design and develop models from literature recommendations

Study by Oguntimilehin *et al*[40] aimed to create a classifier model that can identify Typhoid fever diseases using Rough Set machine learning techniques which is a valuable data analysis and knowledge discovery paradigm that plays a significant role within the broader field of Machine Learning (ML). The researchers worked with a dataset comprising 150 instances. The Rough Set algorithm achieved a training set accuracy of 95% and a testing set accuracy of 96% based on the experiment results. The model was evaluated using a confusion matrix to determine the number of correct and incorrect classifications per class while identifying Typhoid fever disease in unseen data. Although the problem domain is relevant to the one proposed, the researchers did not include diseases that share symptoms with Typhoid fever.

The study by Radhika *et al* [41] developed a classifier model for disease classification based on symptoms using a Decision Tree and Electronic Health Record Analysis. The model was trained to classify 40 symptom-based diseases, including Allergy, GERD, AIDS, Diabetes, Tuberculosis, Asthma, Dengue, Heart attack, Malaria, Typhoid Fever, and more. The Decision Tree initially classifies the disease with the highest confidence level for each of the classified diseases. The author then developed a predictive model to classify symptom-based diseases using the Decision Tree model. The study included foodborne disease like Typhoid Fever which is relevant to the proposed study but the study did not focus on most common foodborne diseases, especially foodborne diseases prevalent in study area. In this study, the researcher might get better accuracy if they use the ensemble machine learning algorithms like RF, XGBOOST and Stacking ensemble learning because those algorithms that can increase the accuracy of the model by boosting the gradient and taking the average of the weak learners performance from the decision tree, and predicting the outcomes[42].

A study by Wang *et al.* [14] used machine learning to develop a model to classify foodborne diseases pathogens. The model used the following features to classify foodborne diseases: location, time of illness, age of patient, type of food consumed, and certain symptoms. The study divided 50,216 samples into training and test sets at a ratio of 7:3. The size of the training set was 35,151 samples, and the size of the test set was 15,065 samples. The study analyzed foodborne disease case data and used machine learning methods such as decision trees, random

forests, k-Nearest Neighbors, stochastic gradient descent, and extremely randomized trees to classify foodborne disease pathogens. The gradient boost decision tree model obtained the highest accuracy, with accuracy approaching 69% in identifying 4 pathogens: Salmonella, Norovirus, Escherichia coli, and Vibrio parahaemolyticus[14] .In this study, the researchers might increase accuracy if they use a stacking ensemble learning method. It leverages the strengths of individual models to potentially achieve superior classification performance[9].

The study Leo *et al* [43] aimed to develop a classifier model that can predict the occurrence of cholera disease by using various machine learning algorithms such as K-Nearest Neighbors (K-NN), Decision Tree, XGBoost, ExtraTree, AdaBoost, Random Forest, and Linear Discriminant Analysis (LDA). The researcher conducted a comparative analysis to identify the best machine learning algorithm that can classify cholera disease. In the study dataset size not explicitly mentioned, but the study deals with imbalanced data. After comparing all the algorithms, the researcher selected XGBoost as the best cholera predictor based on the experiment results. Although this study is relevant to the proposed study, it only focused on cholera disease classification. One of the strengths of this study is that it used ensemble-based algorithms like Random Forest and XGBoost to classify cholera disease.

The study by Grampurohi *et al* [44] aimed to use machine learning algorithms to analyze and classify 41 diseases, with the goal of assisting physicians in diagnosing diseases at an early stage. The paper presents a comprehensive comparative study of three algorithms, namely Decision Tree, Random Forest, and Naïve Bayes, and their performance on a medical record. Each algorithm yielded an accuracy of 95%. The researcher attempted to classify multiple diseases using machine learning algorithms include foodborne disease like Typhoid Fever which is relevant to the proposed study but the study did not focus on most common foodborne diseases, especially foodborne diseases prevalent in the study area.

This study by Bhuiyan *et al* [45] represents a pioneering effort to develop a typhoid fever prediction model that can anticipate outcomes prior to clinical trials. The study utilized a dataset containing 1746 entries with 29 attributes. Both machine learning (ML) and deep learning (DL) approaches are employed and a total of ten algorithms are evaluated. Among these, the XGBoost

classifier emerges as the most promising, achieving an impressive accuracy of 97.87%. This study is relevant to proposed study but it did not include diseases that share symptoms with Typhoid fever. As strong side, this study used ensemble based algorithms like XGBoost to typhoid fever disease.

A study by Kumar *et al.* [46] used artificial intelligence (AI) to develop a model to predict the pathogens in foodborne diseases. They used data from the National Foodborne Disease Surveillance Reporting System on Kaggle, which has 12 attributes. They divided the dataset into training (75%) and testing (25%) datasets. They then compared and contrasted the accuracy of several popular machine learning algorithms for predicting the pathogens in foodborne diseases, including decision trees, random forests, k-nearest neighbors, stochastic gradient descent, and extremely randomized trees. They also developed an ensemble model that combined all of these algorithms. The ensemble model outperformed all of the other classifiers, achieving an average accuracy of 97.26%. The research study is relevant to the proposed research but it only focused on foodborne diseases which are common in India.

Table 2.1 Review of Related Works

Authors	**Title**	**Accuracy and Methodology**	**Drawbacks**
Wang *et al.* [14](2020)	Prediction of Foodborne Diseases Pathogens: A Machine Learning Approach	➢ The authors propose decision trees, random forests, k-Nearest Neighbors, stochastic gradient descent, and extremely randomized trees. ➢ The gradient boost decision tree model obtained the highest	➢ Stacking ensembles could boost accuracy in this study by combining the best of multiple models for stronger classification[9]. ➢ The study did not include foodborne diseases common

		accuracy, with accuracy approaching 69%. ➢ Dataset size 50,216 samples ➢ (7:3 train:test)	in study area like Giardiasis, amoembiasis.
Oguntimilehin *et al*[40](2013)	A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever	➢ Analyzed medical data with Rough Set algorithms. ➢ Dataset size 150 ➢ Achieved 95% accuracy on training data and 96% on unseen test data. ➢ Used confusion matrix to assess correct/incorrect classifications.	➢ Didn't consider diseases with similar symptoms to Typhoid fever. ➢ Ensemble techniques: Considering stacking or other ensemble methods to further improve classification accuracy. ➢ Data was short
Radhika *et al* [41](2020)	Symptoms based disease prediction using decision tree and electronic health record analysis	➢ Used decision tree to classify. ➢ The model could categorize 40 diseases, including common ones.	➢ Foodborne diseases, especially those prevalent in this study area, were not included. ➢ Ensemble machine learning algorithms could improve the model's accuracy.
Bhuiyan *et al* [45] (2023)	Prediction Of Typhoid Using Machine Learning	➢ Both ML and DL approaches are employed,	➢ This study is relevant to proposed study but it did not

	and ANN Prior To Clinical Test,	and a total of ten algorithms are evaluated. ➢ It employed both ML and DL approaches, with XGBoost outperforming other algorithms. ➢ Dataset size 1746	include diseases that share symptoms with Typhoid fever. ➢ As strong side, this study used ensemble based algorithms like XGBoost to typhoid fever disease.

2.5 Summary

Previous research primarily concentrated on classifying generic and chronic diseases, such as heart disease, diabetes, stroke, lung cancer, breast cancer, and malaria. Researchers have employed various machine learning algorithms like SVM, KNN, NB, DT, and RF for disease classification. However, there is a notable gap: no existing machine learning model specifically classifies common foodborne diseases in Ethiopia. In contrast, a study conducted in China developed a machine learning model that classifies foodborne diseases common in that country. Unfortunately, previous research on foodborne disease classification did not include conditions like giardiasis, Amoebiasis both of which are prevalent in Ethiopia especially in study area. This lack of comprehensive classification support poses challenges for healthcare professionals who encounter foodborne diseases with similar symptoms. Consequently, the burden of foodborne diseases in Ethiopia remains high, constituting a significant public health issue. To address this gap, we focused on classifying common foodborne diseases in Ethiopia using machine learning algorithms. The study collected disease data from the specific research area and analyzed the prevalence of foodborne diseases within that region.

Chapter Three

Research Methodology

This chapter deals with the methodology used for the proposed study. Within methodology, the research approaches and methods that are used to collect data and select tools were presented. Also, data analysis and preprocessing are discussed. To come up with solutions the researcher followed the road map (research process) depicted in Figure 3.1 for the proposed research.

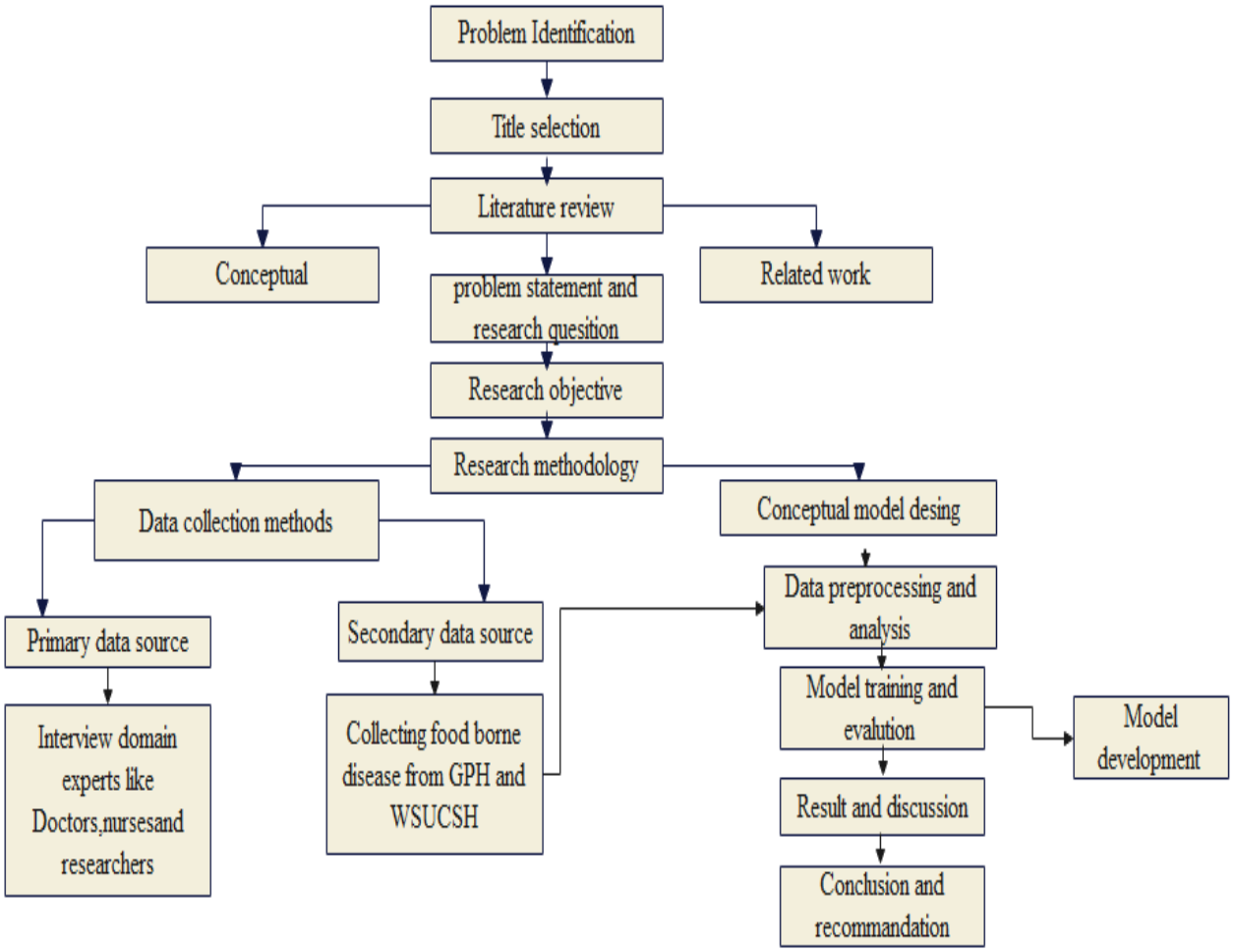


Figure 3.1 Research process of the proposed study

3.1 Research Approach

To reach the study's goal, we used an experimental research design and employed a mixed research approach (both quantitative and qualitative). The quantitative approach is used to collect and analyze numerical data from the study area in order to find patterns and make classifications based on historical data, while the qualitative approach is used to collect and analyze non-numerical data through interviews with domain experts in order to gain a better understanding of the problem domain.

3.2 Research Methods

This research used different techniques and methods to collect important data and to select the tools that are used to design and develop a machine learning model that can classify foodborne disease.

3.2.1 Data Collection Methods

This research collected the data from both primary and secondary sources. From a primary source, the researcher orally interviewed researchers, doctors, and other healthcare professionals in order to understand the problem domain well. To train a machine learning algorithm secondary data is collected from hospitals based on domain expert consultation and the prevalence of the disease in the area. Figure 3.2 illustrates the data collection methods used for the proposed study.

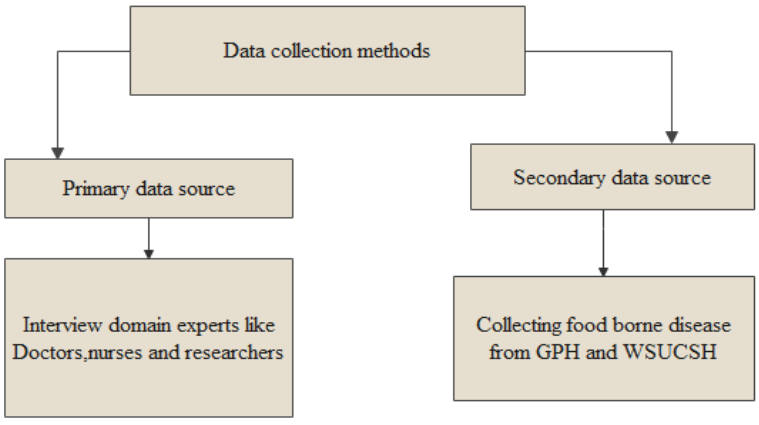


Figure 3.2 Data collection methods for the proposed study

To analyze a large dataset, such as healthcare data, it's difficult to do so and develop a machine learning model since our system will run out of memory while loading those huge datasets. In addition, various constraints such as budget, resources, time, and availability of the data itself constrain the collection of huge amounts of dataset for the proposed study. Thus, purposive sampling techniques were applied in order to permit us to obtain the right information about the population based on data from a subset of the population (sample), rather than having to look into every individual. The researcher followed the sampling framework depicted in Figure 3.3 to collect secondary data from the study area in order to develop the proposed model.

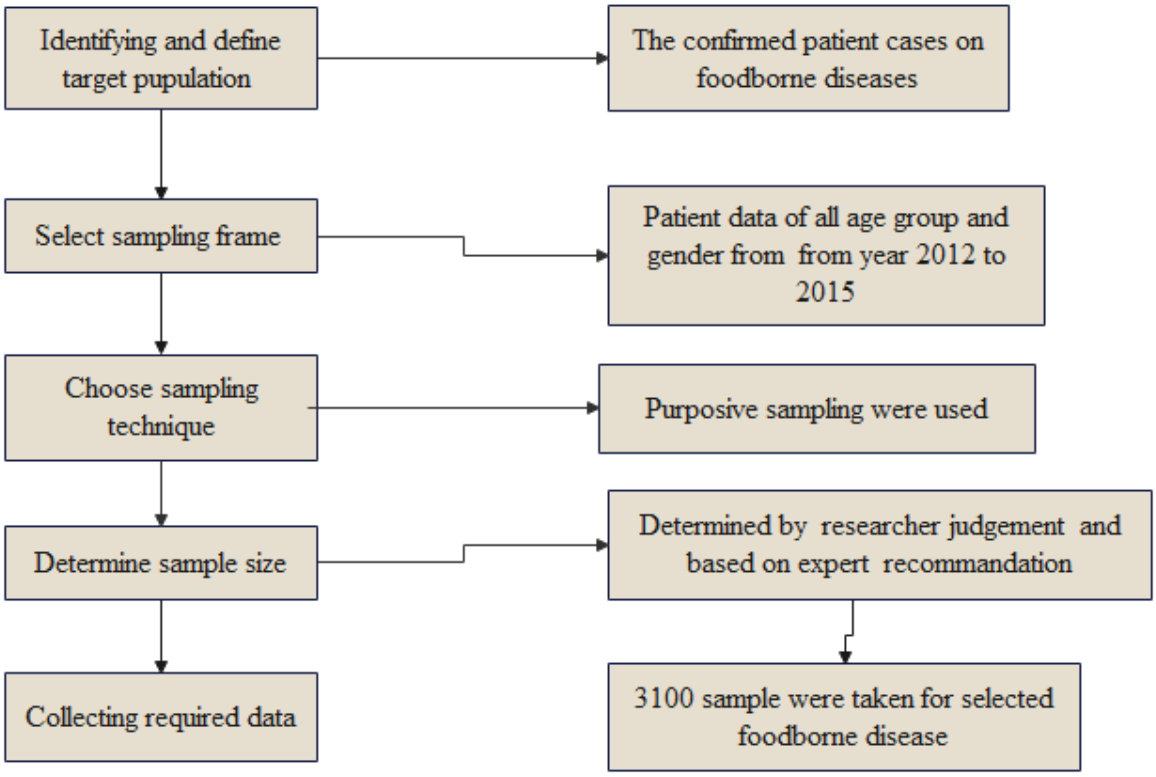


Figure 3.3 Sampling frameworks for data collection.

The confirmed patient cases of foodborne diseases are the target populations that are considered for the study. The different age group and gender that forms the population is included within the sample as a sample frame from the year 2012 to 2015 E.C. To determine the sample size from the confirmed cases, the researcher used the purposive sampling technique based on the recommendation of experts in the field and finally the confirmed cases of foodborne diseases were collected from hospitals.

In this study the dataset was meticulously labeled based on the confirmed diagnoses provided by medical professionals at the hospitals. Additionally, relevant previous research on diseases classification in the area was consulted to ensure consistent and accurate labeling.

3.2.2 Tool Selection Methods

This research has used different tools based on the good features of the tools that are important for the proposed study design. To perform most of the research activities such as data pre-processing, analysis, model training, and testing, python programming and Google colab were used, and to design a conceptual model, E-draw max is used based on their good features such as easy-to-use, large standard and preinstalled library as described in Figure 3.4 below.

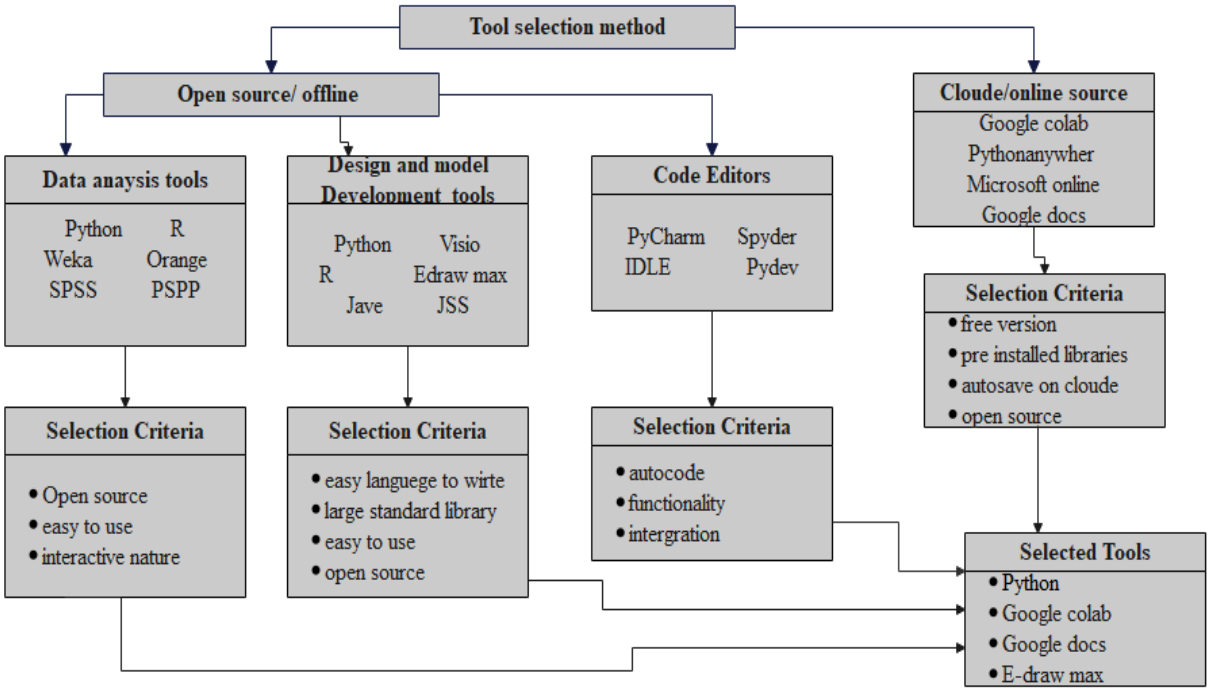


Figure 3.4 Tool selections for the proposed study

3.3 Dataset Description

Secondary data for the proposed study was obtained from two selected hospitals within the Wolaita Zone, southern Ethiopia: Access to data and the opportunity to collaborate with staff at both hospitals significantly facilitated the research process. Therefore, these hospitals were chosen as the primary sources for collecting confirmed cases of foodborne diseases for the proposed study.

The Total number of foodborne diseases data collected from hospitals for model development is 3100 confirmed cases with 24 attributes based on the sampling technique applied. When collecting the data the sample frame that forms the population like age group, gender, and year interval that the sample taken was considered. Table 3.1 below shows the samples gathered for each foodborne disease along with the sample frames, and attributes of each disease (symptoms) as well as, the type of data collected. In order to avoid class imbalance, the researcher determined the year range during data collection. . For instance, the confirmed cases of Typhoid fever in one year (2012 E.C) is greater than the confirmed cases of Giardiasis in three years(2012-2015 E.C) as shown in Table 3.1.

Table 3.1 Dataset Description

No.	Foodborne diseases	No. of sample taken	Age group	Gender	Year interval where the data Taken	Confirmed symptoms(attributes)	Data type
1	Amoebiasis	1060	All	Both	2012-2015 E.C	Diarrhea, Vomiting, Abdominal Cramp, dysentery, Weakness.	Categorical
2	Giardiasis	834	All	Both	2012-2015 E.C	Diarrhea ,Fatigue , Abdominal discomfort, Weight loss, Nausea,	Categorical
3	Typhoid Fever	1206	All	Both	2014 E.C	Fever, chills, headache diarrhea, Abdominal flank, Muscle or joint	Categorical

						pain, cough, Loss of appetite.	
Total = 3100							

The sample dataset before preprocessing is shown in Appendix A.

In addition to the symptoms of foodborne diseases presented in Table 3.1, the description of each selected foodborne diseases are presented here.

Giardiasis is an intestinal parasitic infection caused by the microscopic protozoan Giardia lamblia. It is a common diarrheal disease transmitted through the fecal-oral route. While some infected individuals remain asymptomatic, others experience a range of gastrointestinal symptoms.

Amoebiasis, also known as amoebic dysentery, is an intestinal infection caused by the protozoan parasite Entamoeba histolytica. It primarily affects the colon, causing inflammation and ulceration. Amoebiasis can present as a spectrum of disease, ranging from asymptomatic infection to severe dysentery.

Typhoid fever is a potentially fatal infection caused by the bacterium Salmonella Typhi and it is also contagious and can affect many organs throughout the body.

3.4 Data Analysis

In order to find new patterns, relationships, and insights in the data, we analyzed and visualized the collected data by using various data representation techniques. Figure 3.5 below shows that the confirmed cases of Typhoid fever disease have a high prevalence among others with 1206 confirmed cases within the collected dataset.

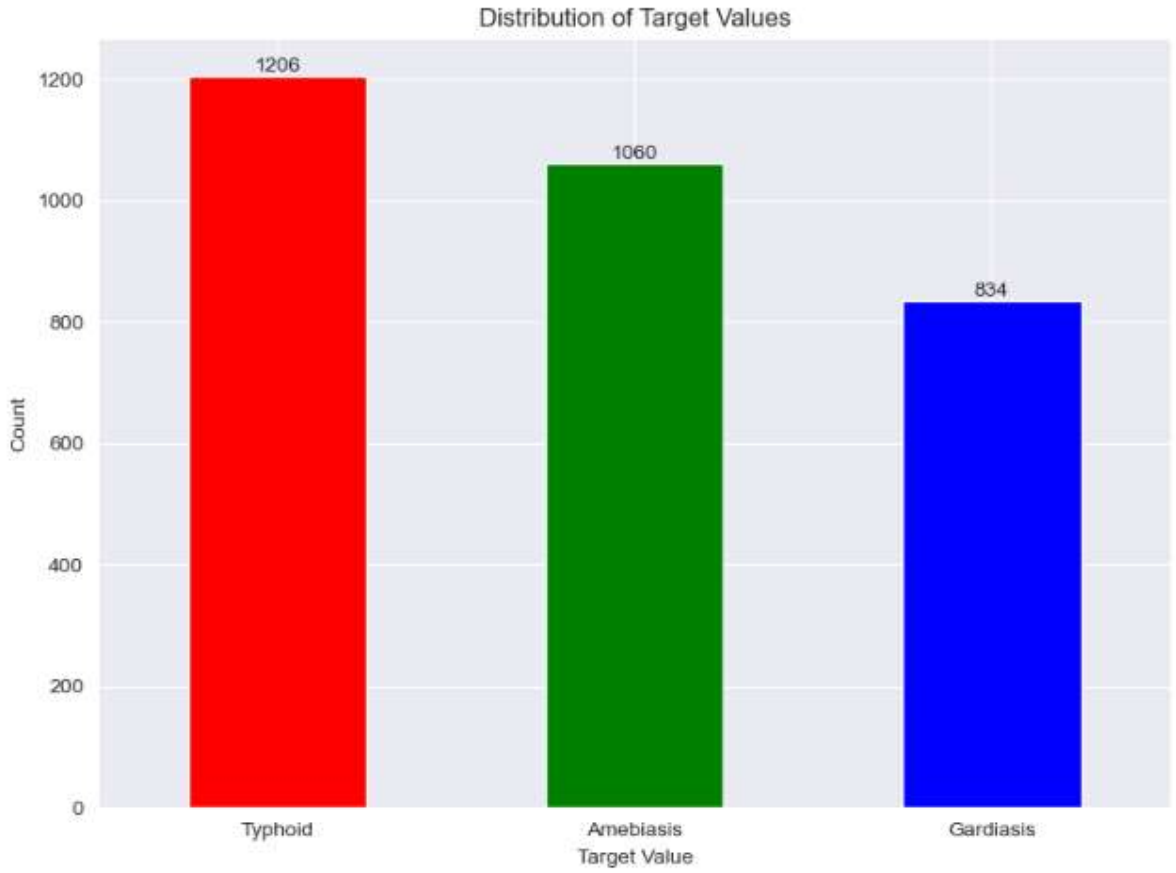


Figure 3.5 The number of confirmed cases for each foodborne disease identified within the collected data.

As we can see from Figure 3.5, the classes are imbalanced. So, to handle this issue we used model based data balancing technique. For instance Random Forest model is suitable model to handle data imbalance by giving maximum weights to minority classes and minimum weights to majority classes while training[8]. In addition, evaluation metrics like precision, recall, and f1_score were used so as to know the generalization capacity of the models in each class.

Most foodborne diseases shares signs and symptoms in common. For instance, Giardiasis and Typhoid Fever have common symptoms like headache, fever, and diarrhea and so on. Thus, it is time-consuming and can make confusion while doctors make a diagnosis of those diseases; however, there is some difference between them. So, to minimize this confusion, promising machine learning algorithms were used in the proposed study to identify or classify the foodborne disease easily.

As we can see from Figure 3.6, 47.7% of Amoebiasis confirmed cases shared symptom AD(abdominal discomfort) in common with 18.6% and 33.7% of Typhoid Fever and Giardiasis confirmed cases respectively.

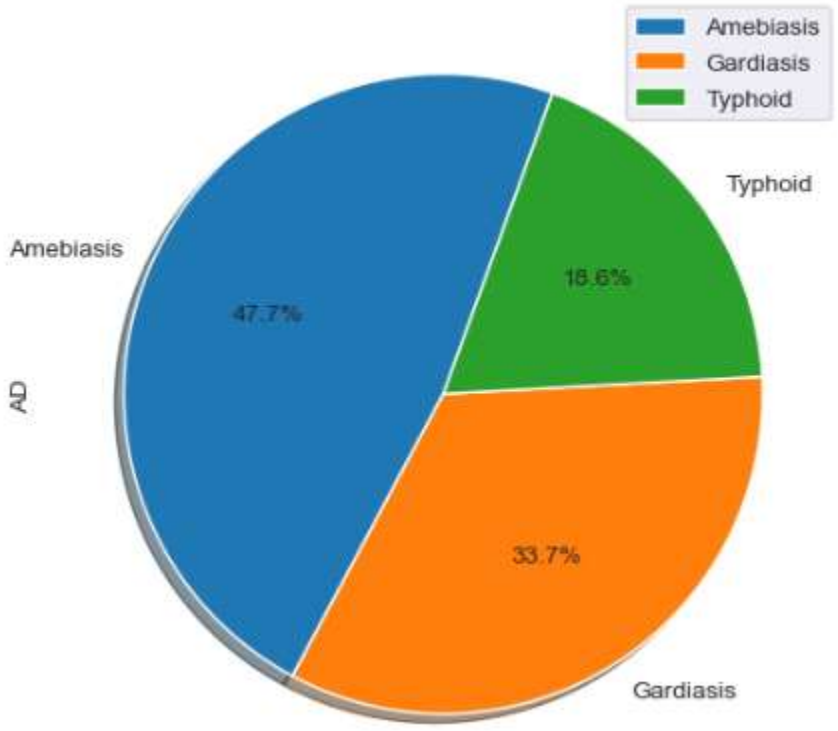


Figure 3.6 Shared symptom (abdominal discomfort) among Amoebiasis, Giardiasis and Typhoid Fever

According to Figure 3.7 below, 54.6% of confirmed cases of Amoebiasis and 30.5 % of confirmed cases of Giardiasis shared the symptom of AF (abdominal flank) in common with 14.9% Typhoid fever confirmed cases.

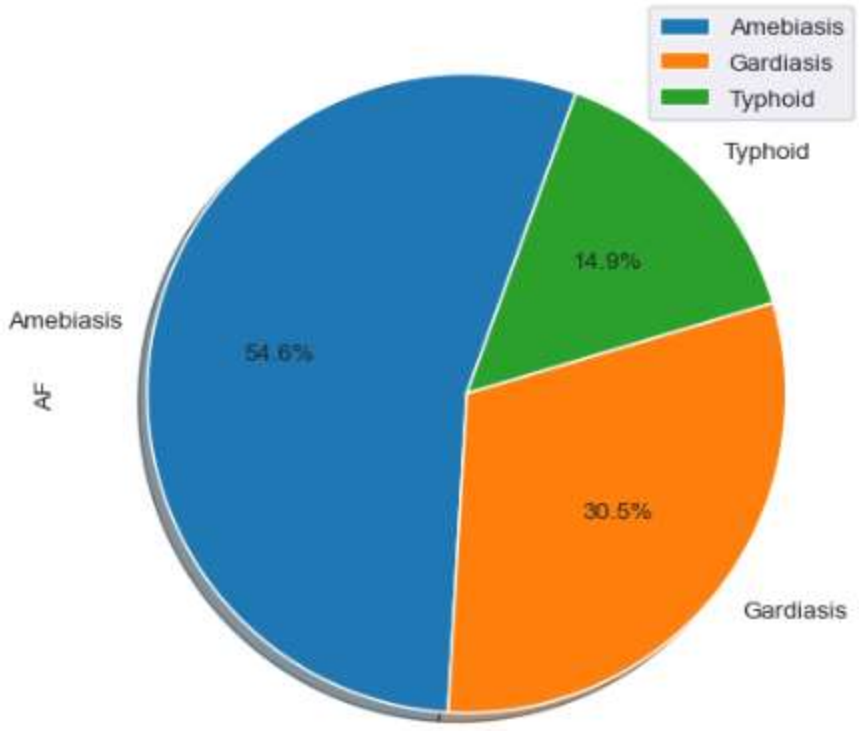


Figure 3.7 Shared symptom (abdominal flank) among Amoebiasis, Giardiasis and Typhoid Fever

As indicated in Figure 3.8 below, 47.2% of confirmed cases of Amoebiasis and 40.0 % of confirmed cases of Giardiasis shared the symptom of diarrhea in common with 12.8% Typhoid fever confirmed cases.

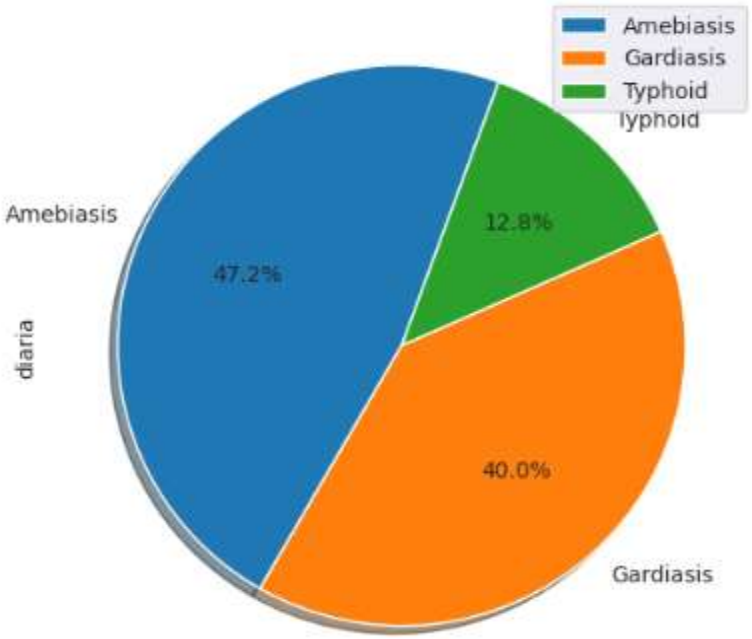


Figure 3.8 Shared symptom (diarrhea) among Amoebiasis, Giardiasis and Typhoid Fever

As shown in Figure 3.9 below, the symptom of loss of appetite is confirmed in 47.7 % of Typhoid Fever confirmed cases, 29.3% Amoebiasis and 23.0% confirmed cases Giardiasis.

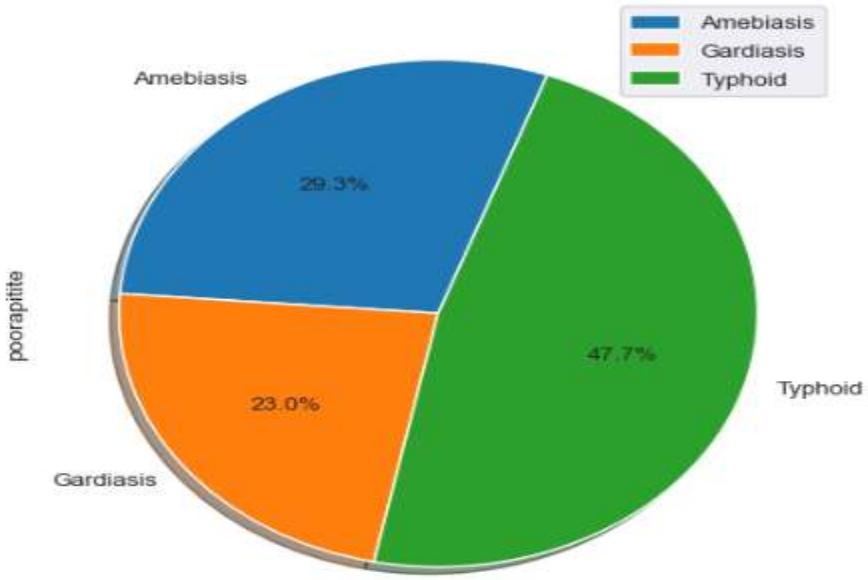


Figure 3.9 Shared symptom (loss of appetite) among Amoebiasis, Giardiasis and Typhoid Fever

As we observe in Figure 3.10, there is a high prevalence of Amoebiasis in children aged 0-10, with 319 cases. This is followed by a decrease in prevalence in subsequent age groups. The prevalence of typhoid fever is highest in young adults aged 21-30, with 352 cases. There is also a relatively high prevalence of typhoid fever in children aged 11-20 (182 cases).Giardiasis prevalence appears to be moderate across all age groups, with the highest prevalence in children aged 31-40 (194 cases).

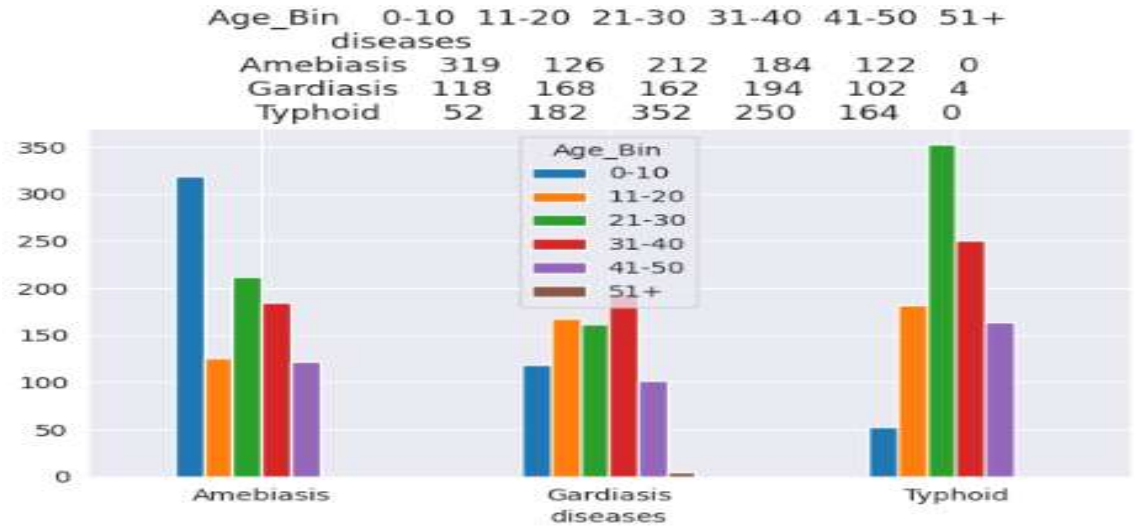


Figure 3.10 Prevalence of each disease in different age group

As we can see from Figure 3.11 all three diseases, amoebiasis, giardiasis, and typhoid fever, appear to have a similar number of cases overall. However, there are some differences in the male-female distribution of cases. For both amoebiasis and giardiasis, there are more male cases than female cases. amoebiasis has the most male cases (636) compared to females (424). Giardiasis also has more male cases (496) than female cases (338). Typhoid fever is the only disease in the image that shows more female cases (642) than male cases (564).

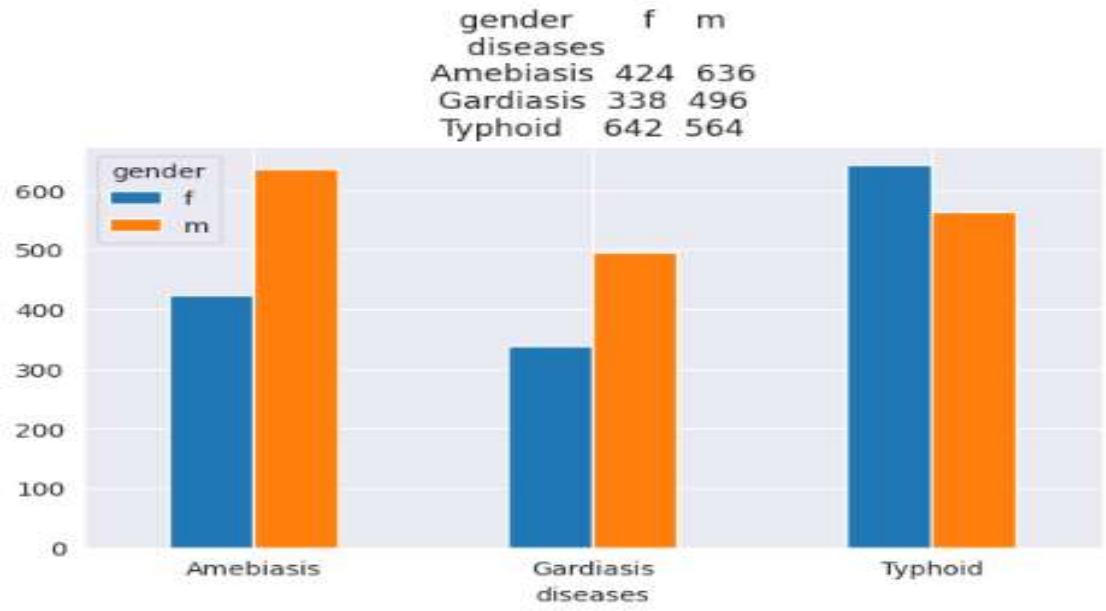


Figure 3.11 Prevalence of each disease in gender

3.5 Data Preprocessing

Before training the model, the data should be preprocessed in such a way that a model can understand. Preprocessing activities such as data standardization and transformation, finding missing values from a dataset, feature selection based on its importance and data training and test splitting were conducted in order to train a model with quality data.

3.5.1 Data Standardization and Transformation

Before training a model, the data were standardized at the same scale in order to prevent model over fitting. In our case, the dataset already has the same scale or standard, so that, there is no need of standardization, we opted to standardize the 'age' column specifically. All data type in a dataset is changed to numeric, because, we represent the data to 1 if the symptom in a specific disease is present and 0 if the symptoms are not present i.e., all columns that contain symptoms are filled with 1 if the symptom occurred in specific disease and 0 if the symptom does not occur. In this study, researchers used the min-max normalization technique to normalize the age attribute (column). The formula of min-max normalization is shown below in Equation 3.1.

$$X' = \frac{X - min(X)}{max(X) - min(X)} \quad (3.1)$$

Where X is an original value, and X' is the normalized values, min(X) is the minimum value of the original dataset, and max(X) is the maximum value of the original dataset.

In addition, some of the column with data type categorical like gender was transformed to the numeric value in order to train a model. To do so, one hot encoder is used to transform. It is a powerful technique in machine learning for transforming nominal categorical data into a numeric format. It preserves independence by creating new binary variables for each unique category, ensuring no artificial ordinal relationship is imposed. This approach avoids misinterpretation, as label encoding assigns integer values based on alphabetical order, which can introduce misleading rankings[47]. So the "gender" column would be split into two new columns: "M" and "F," with 1 indicating the presence of that gender and 0 otherwise. In the dataset, the target variable "Disease Type" contains three nominal values: Typhoid Fever, Giardiasis, and Amoebiasis. In order to transform target variable into numerical form a label encoder is used to transform labels with data type nominal to numeric because it is a popular technique to convert nominal data into numeric. In this method, the unique number is given to each label based on alphabetical order. So, the disease type (target class) column such as Amoebiasis , Giardiasis, and Typhoid Fever is transformed to 0,1 and 2 respectively as shown in Table 3.2. The dataset after preprocessing is shown in Appendix B.

Table 3.2 Data Transformation

Disease type	Transformed value	Male		Female		Symptoms	
		Present	Absent	Present	Absent	+ve	-ve
Amoebiasis	0	1	0	1	0	1	0
Giardiasis	1	1	0	1	0	1	0
Typhoid Fever	2	1	0	1	0	1	0

3.5.2 Finding Missing Value

When we train a model with data that contains a missing value, it may generate an error and the model is also unable to train it. So finding the missing value and filling it with the right data is required to train a model and to have a good performance on it. In our case, 5 missing values were found and the missing value is imputed using the mode method which is a preferable method for categorical numeric data imputation. It handles missing values by replacing them with the most frequent value of the observed feature[48].

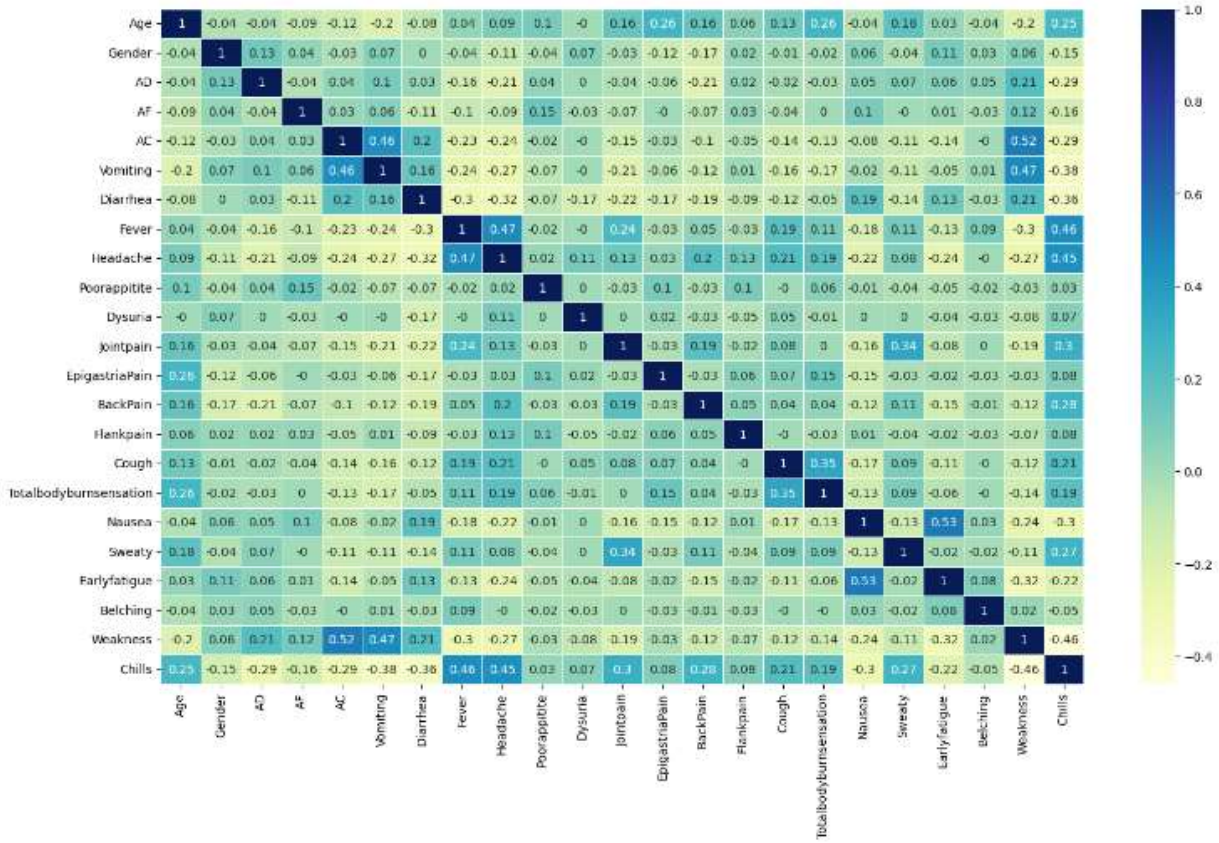
3.5.3 Correlation Matrix

To know the relation between two variables and to determine the factor of change of one variable on another variable within the collected dataset, the correlation coefficient is computed. In addition to this, the correlation matrix is used as input for other analyses, like exploratory factor analysis (used to know the important features on the dataset). The correlation coefficient value falls between -1 to +1. We can interpret its values as follow:

According to a Pearson correlation coefficient[49], if the coefficient value is 0, it indicates that the two variables don't have a linear relation. Positive /negatively weekly correlation exists between the two variables if it is between 0 and 0.3 (0 and -0.3). Additionally, if it falls between 0.3 and 0.7 (-0.3 to -0.7), the two variables have a moderately positive/negative correlation. And the relationship between two variables is strongly positive/negative if it is between 0.7 and 1 (-0.7 and -1). Notably, each variable is perfectly associated with itself, as seen by the correlation coefficients along the table's diagonal being equal to 1. No interpretation can be made from these cells.

As we can see from Table 3.3 the correlation between two particular variables is displayed in each cell. For instance, nausea and early fatigue have a correlation of 0.53, indicating the two variables have a moderately positive association i.e. “nausea” is moderately correlated with more “early fatigue”. And the “belching” and “weakness” have a correlation of 0.01, which denotes a weakly negative relation. And cough with joint pain, joint pain with dysuria have a correlation of 0, which indicates that the two variables don't have a linear relation.

Table 3.3 Correlation matrix for each attribute in a dataset



3.5.4 Feature Importance

Feature importance refers to methods that rate input features according to how well they can predict a target variable. The feature importance score is the basis for dimensionality reduction and feature selection that can increase the efficiency and effectiveness of a predictive model on the problem. To do so, random forest feature importance analysis is utilized due to its popularity to know the importance of each feature[50]. It determines the importance of each feature by improving the purity of the leaves used by each tree in the random forest. Where, the importance of the features increases as the increase in leaf purity increases.

Table 3.4 Feature importance of each attribute in dataset

No.	Feature's name	Score of feature
1	Weakness	0.243319
2	Chills	0.196843
3	Nausea	0.120400
4	early fatigue	0.088068
5	Age	0.058681
6	AC(abdominal cramp)	0.044468
7	Headache	0.044365
8	Diarrhea	0.039404
9	Vomiting	0.028162
10	AD(abdominal discomfort)	0.023372
11	Fever	0.022850
12	Back Pain	0.016247
13	Joint pain	0.013212
14	Gender	0.011061
15	Poor appetite	0.007643
16	Total body burn sensation	0.006271
17	Epigastria Pain	0.006240
18	Cough	0.005984
19	Sweaty	0.005933
20	AF(abdominal Flank)	0.005789
21	Flank pain	0.005487
22	Dysuria	0.004345
23	Belching	0.001857

As, indicated in Table 3.4, the first 4 features namely weakness, chills, nausea and early fatigue in the table are the most important features to predict the target classes. Based on random forest feature importance analysis some of the features like flank pain, dysuria and belching rated less score and they are subjected to be dropped. However, according to domain expert consultation,

all features have their own shares while making foodborne disease diagnosis. So that, there is no need of dropping features. Thus, the researcher used all features for foodborne disease classification. To answer the question raised under research question 1, the researcher gathered both primary and secondary data of foodborne diseases, doing feature analysis on collected secondary data using random forest feature importance and domain expert consultation. In addition to this, the correlation coefficient matrix for each attribute is computed to know the relationship between each feature. By doing so, the researcher identified the importance of each feature with in a dataset for foodborne diseases classification. Therefore, the question raised by the researcher under research question 1 is answered here.

3.5.5 Train and Test split

To train a model, the dataset was split into a training set and a test set because to evaluate the model, an unseen data is required to check whether the model performs well or not. 80/20 and 70/30 split is applied for the proposed model development based on their popularity and a suggestion of previous researchers [51].

Chapter Four

Proposed Model Design and Development

The proposed model design and development, the models selected for development as well as the metrics used to evaluate the performance of the models are discussed here.

4.1 Proposed Model Design of Foodborne Disease Classification

An experimental research design is applied to attain the proposed study's goal because a suitable algorithm is chosen based on the experiments conducted on each selected machine learning model.

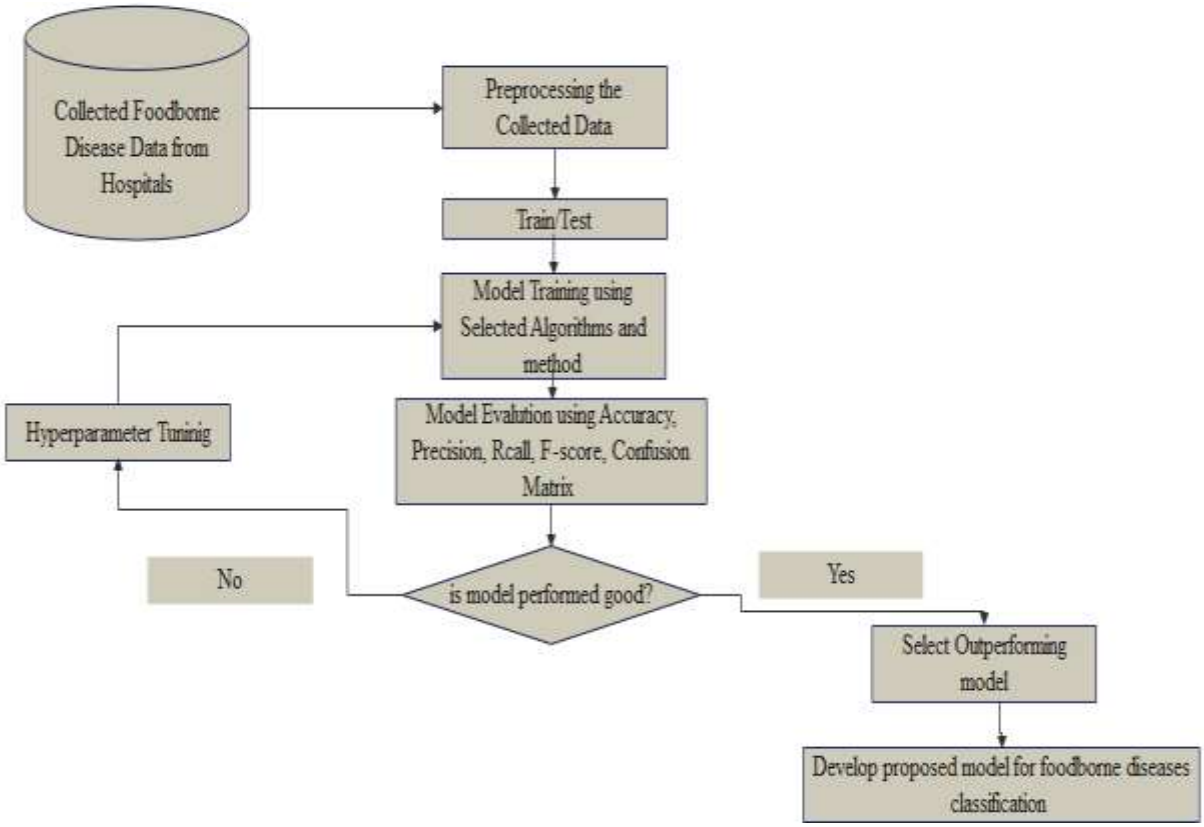


Figure 4.1 Proposed research design of foodborne disease classification

Figure 4.1 above illustrates that, to develop a proposed model, first, the data on the most prevalent foodborne diseases were gathered by the researcher from hospitals. In order to make the datasets suitable for the chosen machine learning model, the datasets were preprocessed. The state of art machine learning algorithms was chosen to train a model after preprocessing, and its performance is then evaluated using data that has not yet been seen (hold-out or test set). Additionally, this research implemented a stacking ensemble method. Finally, the suitable and outperforming model is selected among others and the proposed model is developed.

4.2 Proposed Model Development

To train and develop a machine learning models, the previous studies used popular and state- of-the-art machine learning algorithms such as Decision Tree, Random Forest and XGBoost for classification problem[28], [52]. To achieve optimal model performance, we trained models utilizing these algorithms alongside a stacking ensemble method. Through a rigorous process of experimentation and comparative analysis, the most suitable model selected for further development as the proposed solution for foodborne disease classification.

4.2.1 Decision Tree Model

We chose a decision tree to classify foodborne illnesses because it's easy to understand how it works. This method asks broad symptom questions (like fever) at first and splits the data based on the answer (yes or no). Then, it gets more specific with each branch, using additional symptom questions to tell different diseases apart. This goes on until it reaches a final diagnosis at the end (leaf node). However, decision trees can learn unimportant details from the training data, so adjusting settings (hyperparameter tuning) is essential to control the tree's complexity.

There are several parameters to consider when training a decision tree model in Python. These parameters can affect how well the model performs, including how well data is separated (criterion), maximum tree depth (max_depth), number of features considered at each split (max_features), minimum data points at leaf nodes (min_samples_leaf), and minimum data points for splitting branches (min_samples_split)[53]. We used a grid search approach to find the best settings for their model, which helps identify the combination of settings that leads to the most accurate predictions.

4.2.2 Random Forest Model

We tried a random forest algorithm so as to have better accuracy, hence, the random forest algorithm takes all the decision made by the decision tree and takes an average of it, and make a classification. When classifying the foodborne diseases using a random forest algorithm, it takes an average of the diseases classified by the decision tree and classifies the specific disease or it can classify the diseases based on the majority vote of the decision tree i.e. if the majority of decision tree classifies one disease, so, the algorithm votes or selects that disease as target class. A random forest classifier is a multi-tree ensemble model and there are various parameters to be considered when training the RF model in Python. All of which might influence the model's performance as (n_ estimator, max_ features, min_ samples leaf, criterion, and max_ depth)[54]. Where the number of decision trees in the forest is represented by n estimators, the maximum number of features that can be separated in a node is max features, the number of levels in the tree that can be reached is max depth and when selecting whether or not to split the tree, we must consider criteria such as entropy, gini, and others. The above-mentioned parameter values were selected based on grid search algorithm while training a model, because it is used to find the optimal hyper-parameters of a model with the most accurate predictions [55].

4.2.3 eXtreme Gradient Boosting Model

We experimented with the XGBoost algorithm which boosts weak learners from decision trees so as to have better accuracy of a model. To do foodborne diseases classification using the XGBoost algorithm, first, the algorithm looks at each decision tree classification loss function and if it is high, it boosts the gradient so as to minimize the loss function and to have better model performance. Finally, it takes the summation of all decision tree loss functions and predicts the output or target class (foodborne disease). The most crucial hyper-parameters for XGBoost are those that have the best chance of enabling the algorithm to produce the most accurate, unbiased results and the quickest without over-fitting: How many sub-trees to train, the maximum tree depth (a regularization hyper-parameter), the learning rate, the L1 (reg alpha) and L2 (reg lambda) regularization rates that determine the extremeness of weights on the leaves, the complexity control (gamma), a pseudo-regularization hyper-parameter and the minimum child weight are considered[56].

4.2.4 Stacking ensemble method

We investigated a stacking ensemble learning approach that leverages the strengths of multiple machine learning algorithms. This approach utilizes a combination of base learners, each trained independently, and a meta-learner that integrates their predictions for a more robust and accurate classification. This study employed three base learners in this ensemble: XGBoost, Decision Tree, and Random Forest. Each learner offers distinct advantages. Decision trees are interpretable, allowing us to understand the decision-making process. Random forests improve stability by averaging predictions from multiple decision trees. XGBoost excels in handling complex relationships within the data. The researcher employed the trained base learners on the prepared foodborne disease dataset. Each learner independently analyzed the data and generated its own predictions for the disease categories. These initial predictions served as the foundation for the stacking approach. A logistic regression model acted as the meta-learner in this ensemble. It is a popular choice for the meta-classifier due to its interpretability, computational efficiency, and ability to provide probabilistic interpretations of class membership[57]. The outputs, or predictions, from the base learners (XGBoost, Decision Tree, Random Forest) were fed as new features into the logistic regression model. The actual disease labels in the dataset remained the target variable. Essentially, the meta-learner was trained to learn how to best combine the often-complementary predictions from the base learners, aiming to create a more accurate final prediction. Once trained, the logistic regression meta-learner was utilized to generate the final classifications for foodborne diseases. By leveraging the collective knowledge of the tuned base learners, the stacking ensemble aimed to surpass the performance of any single base learner in classifying foodborne diseases. This approach not only benefits from the strengths of individual learners but also mitigates potential weaknesses of any single algorithm. The hyper parameter tuning for each base learner further enhances the overall performance and generalizability of the stacking ensemble for foodborne disease classification.

4.3 Proposed Model Evaluation

After training the models we evaluated the performance of them by using different machine learning evaluation metrics such as classification accuracy, confusion matrix, precision, recall, and f1_ score which were also used in previous studies[33].

4.3.1 Confusion Matrix

It evaluates the classification models' performance on test data and determines how good our classification model is, when classifying foodborne diseases.

There are 4 important terms in the confusion matrix as mentioned in Table 4.1:

True Positives: When the predicted output is YES and the actual outcome is also YES.

True Negatives: These are the occasions where we predicted NO and the actual result is also NO.

False Positives: These are instances where we predicted YES but the actual result was NO.

False Negatives: These are occasions where we expected NO but got YES instead.

For instance, in our case, the random forest model in Amoebiasis disease, classified 205 instances as true positive, 2 instances as false positive, and 7 instances as false negative i.e. out of 214 confirmed cases 9 cases are wrongly classified as shown in Table 5.7.

Table 4.1 Components of confusion matrix

		Predicted classification		
Actual classification	Class	Amoebiasis	Giardiasis	Typhoid Fever
	Amoebiasis	True Amoebiasis	False Giardiasis	False Typhoid Fever
	Giardiasis	False Amoebiasis	True Giardiasis	False Typhoid Fever
	Typhoid Fever	False Amoebiasis	False Giardiasis	True Typhoid Fever

4.3.2 Classification Accuracy

Classification accuracy is a metric that measures a classification model's performance by dividing the number of correct predictions by the summation of all predictions made as shown in Equation 4.1 [58]

$$accuracy = \frac{number\ of\ correct\ prediction}{Total\ number\ of\ predction} = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(}4.1\textbf{)}$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

4.3.3 Precision

Precision refers, to how many of the positive classes predicted accurately by the model were actually true. It is formulated as in Equation 4.2

$$\text{precision} = \frac{TP}{TP+FP} \quad (4.2)$$

4.3.4 Recall

It is defined as the percentage of positive classes that our model accurately predicted out of a total of positive classes as shown in Equation 4.3

$$\textbf{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TPi+FN}}$$

(4.3)

4.3.5 F1-Score

The F1-score integrates a classifier's precision and recall into a single metric by calculating their harmonic mean; it is typically used to evaluate the performance of classifiers as shown in

$$\text{F1} - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (4.4)$$

4.4 Machine Learning Libraries used for Proposed Model Development

To make Python programming easier and faster for proposed model development, a variety of Python libraries were used. We won't have to create the same code for each program; instead, we have just imported the appropriate library for the intended program. Table 4.2 shows the libraries used for the proposed model development.

Table 4.2 Important libraries used in the proposed study

No.	Name of library	Use
1	NumPy	Library used for working with arrays.
2	Pandas	Used to load and analyze data.
3	Sklearn	Used to preprocess the data and develop a machine learning models.
5	Matplotlib	Used to visualize the data

Chapter Five

Experimental Result and Discussion

This chapter presents the experimental setups used to develop the proposed model, the data preparation for model development and the hyper-parameter tuning, and the result obtained from the selected machine learning algorithms such as Decision Tree, Random Forest, XGBoost, and stacking ensemble learning and their performance is evaluated and compared using different evaluation metrics such as confusion matrix, accuracy, precision, recall, and f1 score. Finally, a suitable algorithm is selected to develop the proposed model based on the experiments conducted.

5.1 Experimental Setup

To conduct the experiment, the researcher used a machine that has the following qualifications:

- Laptop computer with Intel(R) Core (TM) i5 -4210M
- CPU with 2.6 GHZ speed
- 8GB RAM
- 250 GB SSD storage
- Window 11
- GPU is also used while running python code in Google Colab.

5.2 Dataset Preparation for Model Development

After collecting the necessary data, preparing the data for model development is a critical step in producing a high-performing model. To develop the proposed model, the confirmed cases of foodborne diseases data were collected and prepared in such a way that the model can understand. The total dataset which is prepared to build a proposed model is 3100 instances with 24 attributes. Among those data, 2170 were reserved for training, 930 data were kept for testing based on 70/30 splitting and 2480 data were used in the training set and 620 data were used in the testing set in the 80/20 split. Figure 5.1 below shows the number of each foodborne disease data in percentage within our dataset.

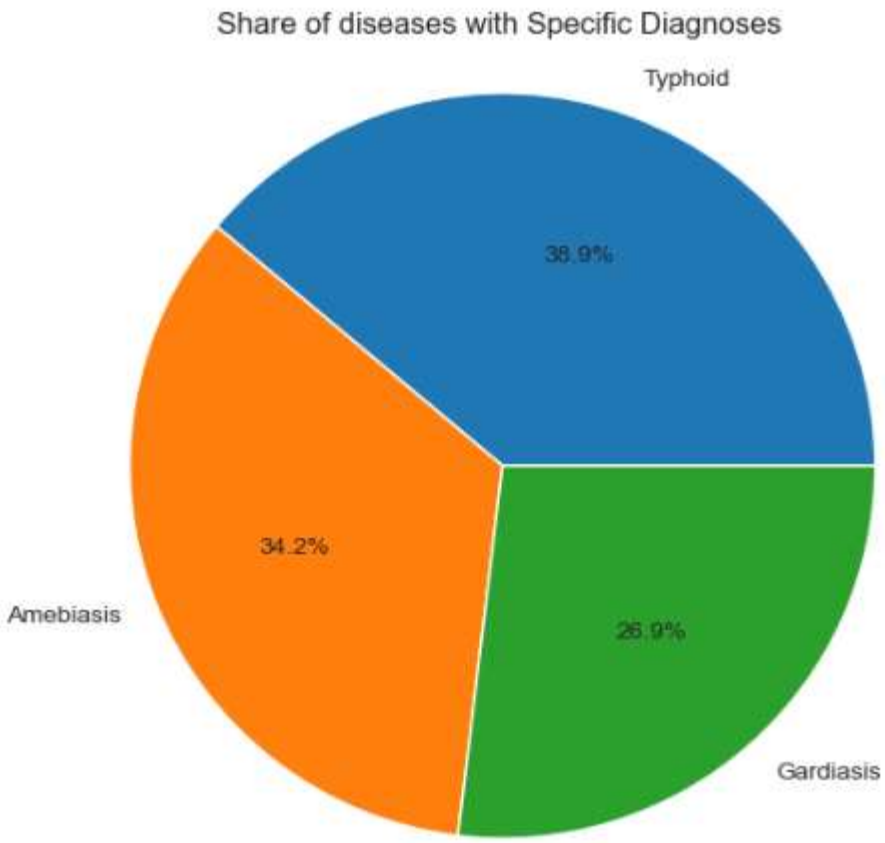


Figure 5.1 Count of foodborne diseases in percent within a dataset

5.3 Machine Learning Model Selection

After preparing the data, ML models were chosen, trained, and tested throughout the modeling process. We tried a state of art machine learning algorithms such as Decision Tree, Random Forests, XGBoost, and Stacking ensemble learning model and compared their performance to build a good model. Finally, the outperforming model is selected to develop a proposed model based on the experiment conducted.

5.4 Hyper-parameter Tuning

Once we have constructed and tested our model, we examined whether its accuracy may be improved in any way. This is done by tuning the parameters for each model. The parameters contained in each model were selected in order to improve the accuracy of a specific model. To find these settings, parameter tuning is applied for all models by using grid search algorithms. The following Table 5.1 shows the parameters and their search space for each model and their

corresponding values to train the model using a grid search algorithm. Taking all search space and training the model with all of them is very hard because it takes long time to execute all fits. Thus, the researcher does not fully addressed all of the search spaces while training the model, instead taking the one which is recommended by the previous researchers [59] and used trial and error method to select the best parameters.

Table 5.1 Hyper-Parameter Tuning for all models

No.	Model	Parameters	Search space	Selected Parameter Values
1	Decision Tree	Criterion	'gini', 'entropy'	'gini'
		max_depth	20,21,23,24,25,26,None	23
		max_features	'int', 'float', 'sqrt','log2',None	None
		min_samples_leaf	1,3,5	1
		min_samples_split	2,5,10	2
2	XGBOOST	learning_rate	0.001,0.01,0.1	0.1
		max_depth	3,4,5,6,8,10,12,15	15
		min_child_weight	1,3,5,7	1
		Gamma	0.0,0.1,0.2,0.3,0.4	0.0
		colsample_by tree	0.3, 0.4, 0.5 , 0.7	0.5
3	Random forest	Criterion	'gini', 'entropy'	Gini
		max_depth	5, 10, 15,20,24,None	15
		max_features	'sqrt', 'log2', None	Sqrt
		min_samples_leaf	1, 2, 4	1
		min_samples_split,	2, 5, 10	2
		n_estimators	100,150, 200, 300	200

5.5 Experiment Results

In this section, the performance of each selected machine learning model is evaluated using different evaluation metrics such as accuracy, precision, recall, f1-score, and confusion matrix.

5.5.1 Decision Tree

In order to enhance the performance of the decision tree model, numerous parameters were employed, including criterion, max_depth, max_feature, min_samples_leaf, min_samples_split as stated in Table 5.1. Decision tree is a non-parametric learner known for interpretability but prone to overfitting[41]. Therefore, various evaluation metrics, including precision, recall, and fl-score, are used to evaluate the decision tree model's performance in each class in both 70/30 and 80/20 train test splits as shown in Table 5.2.

Table 5.2 DT model evaluation on each class using precision, recall, and f1_score

Classes	Train test Split	Precision	Recall	F1 Score
Amoebiasis	70/30	0.96	0.97	0.96
	80/20	0.95	0.98	0.96
Giardiasis	70/30	0.96	0.93	0.95
	80/20	0.96	0.95	0.95
Typhoid Fever	70/30	0.96	0.98	0.97
	80/20	0.97	0.96	0.96
Macro avg	70/30	0.96	0.96	0.99
	80/20	0.96	0.96	0.96
Weighted avg	70/30	0.96	0.96	0.96
	80/20	0.96	0.96	0.96

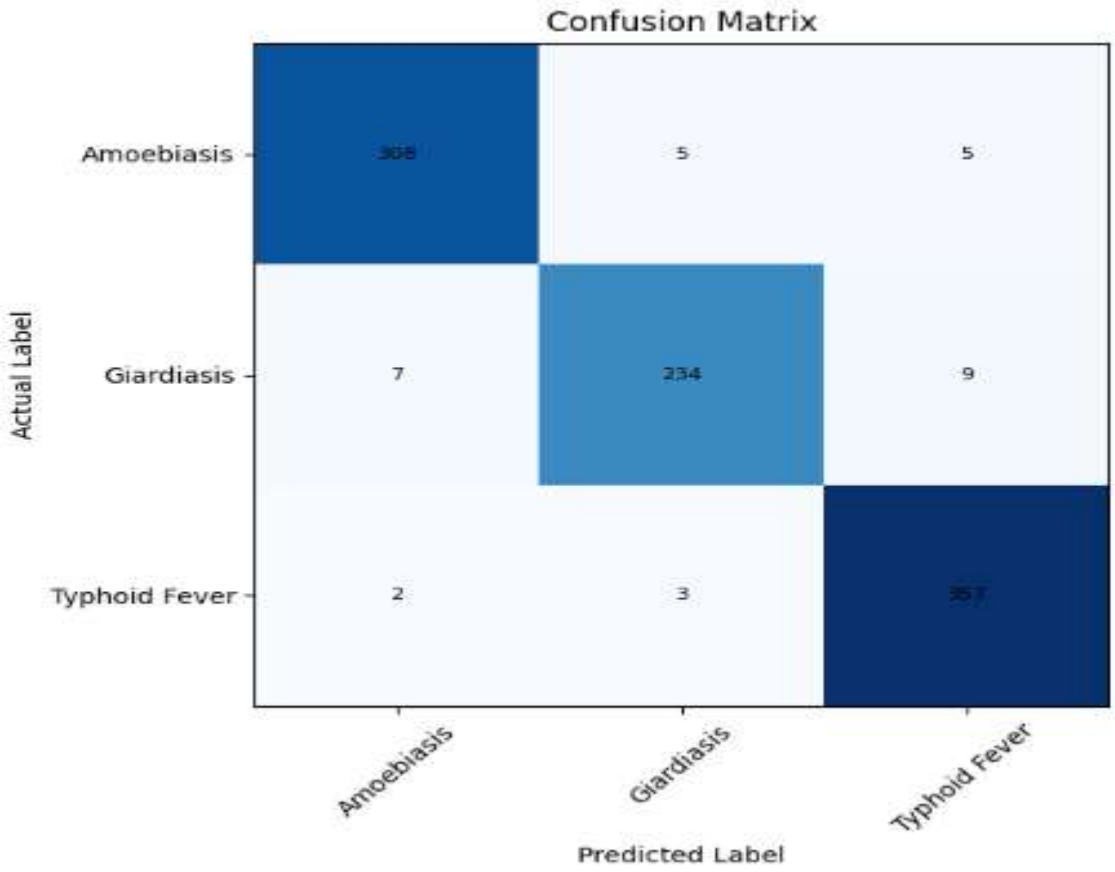
The train and test accuracy of the Decision Tree model on foodborne disease classification was carried out in order to examine how well the model was trained and classified the foodborne diseases using unseen data. According to the experiment results obtained from the decision tree model, the training accuracy is 99.45% and the test accuracy is 96.67% in a 70/30 split and 99.44% of train and 96.45% of the test were obtained from 80/20 and also the overall performance of random forest model in terms of precision, recall and f1_score were depicted below in Table 5.3.

Table 5.3 Evaluation of DT model in accuracy, precision, recall, and f1_score

Model Name	Train test Split	Precision	Recall	Fl_score	Train Accuracy	Test accuracy
DT	70/30	96.67%	96.67%	96.66%	99.45%	96.67%
	80/20	96.47%	96.45%	96.45%	99.44%	96.45%

In the confusion matrix, which is depicted in Table 5.4 below, the findings of the Decision Tree model for the classification of foodborne diseases are enumerated. The decision tree model classified 899 instances correctly out of 930 instances and incorrectly classified 31 instances. The interpretation of each classes are discussed as follow: The decision tree model correctly classified 308 Amoebiasis instances of and incorrectly classified 10 instances (wrongly classified 5 instances as cases of Giardiasis and 5 instances as cases of Typhoid Fever), while the decision tree model correctly classified all 234 instances of Giardiasis and wrongly classified 16 instances (wrongly classified 7 instances as Amoebiasis class and 9 instances as Typhoid Fever class), in Typhoid Fever class 357 instances are correctly classified and wrongly classified 5 instances (wrongly classified 2 instances as Amoebiasis class and 5 instances as Giardiasis class). As per the result shown in Table 5.4, we can say Amoebiasis, Giardiasis and Typhoid Fever disease have shared symptoms in common.

Table 5.4 Confusion matrix of DT model



5.5.2 Random Forest

In order to enhance the performance of the random forest model, numerous parameters were employed including criterion,max_depth,max_features,min_samples_leaf,min_samples_split,n_estimators as stated in Table 5.1. Therefore, various evaluation metrics, including precision, recall, and fl score, are used to evaluate the random forest model's performance in each class in both 70/30 and 80/20 train test splits as shown in Table 5.5.

Table 5.5 RF model evaluation on each class using precision, recall, and f1_score

Classes	Train test Split	Precision	Recall	F1 Score
Amoebiasis	70/30	0.99	0.96	0.97
	80/20	0.99	0.97	0.98
Giardiasis	70/30	0.96	0.95	0.96
	80/20	0.98	0.95	0.96
Typhoid Fever	70/30	0.96	1.00	0.98
	80/20	0.96	1.00	0.98
Macro avg	70/30	0.97	0.97	0.97
	80/20	0.98	0.97	0.97
Weighted avg	70/30	0.97	0.97	0.97
	80/20	0.97	0.97	0.97

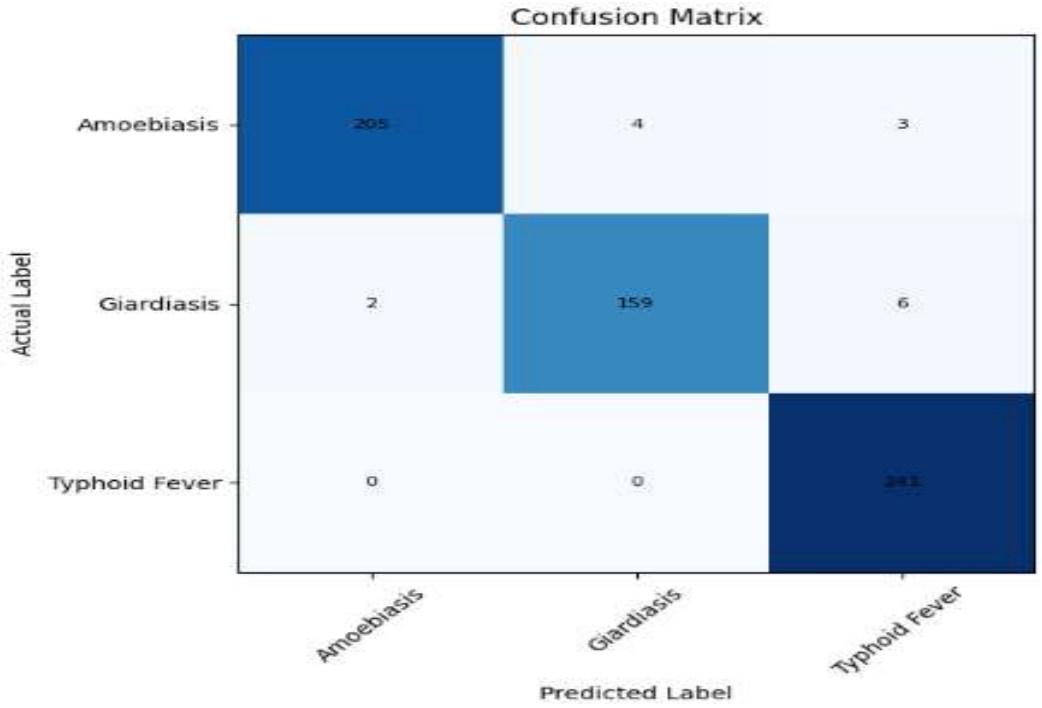
The train and test accuracy of the random forest model on foodborne disease classification was carried out in order to examine how well the model was trained and classified the foodborne diseases using unseen data. According to the experiment results obtained from the random forest model, the training accuracy is 99.45% and the test accuracy is 97.10% in a 70/30 split and 99.44% of train and 97.58% of the test were obtained from 80/20 and also the overall performance of random forest model in terms of precision, recall and f1_score were depicted below in Table 5.6.

Table 5.6 Evaluation of RF model in accuracy, precision, recall, and f1_score

Model Name	Train test Split	Precision	Recall	Fl_score	Train Accuracy	Test accuracy
RF	70/30	97.12%	97.10%	97.08%	99.45%	97.10%
	80/20	97.61%	97.58%	97.57%	99.44%	97.58%

The results of the random forest model on foodborne disease classification are summarized in the confusion matrix to show the number of correct and wrong classifications per class in 80/20 split. The Random Forest model correctly classified 605 instances and incorrectly classified 15 instances out of 620 instances which were taken as a test set. Here, the detailed interpretation of each classes are described as follow: in the Amoebiasis class, the RF model correctly classified 205 instances and wrongly classified 7 instances (wrongly classified 3 instances as Giardiasis class and 4 instances as Typhoid Fever class), and the random forest model correctly classified all 159 instances in the Giardiasis class and wrongly classified 8 instances (wrongly classified 2 instances as Amoebiasis class and 6 instances as Typhoid Fever class), in Typhoid Fever class 241 instances are correctly classified and wrongly classified 0 instances. Based on the confusion matrix result depicted in Table 5.7, the researcher believes that the disease Amobiasis, Giardiasis and Typhoid Fever have shared the symptoms in common as indicated in Figure 3.6 -3.9.

Table 5.7 Confusion matrix of RF model



5.5.3 eXtreme Gradient Boosting

In the XGBoost model five parameters such as colsample_bytree, gamma, learning_rate, max_depth, min_child_weight were employed to increase the performance of the model. The grid search algorithm's suggestions were used to determine parameter values. The parameters utilized in the XGBoost model and their selected values based on the grid search technique are shown in Table 5.1.

The XGBoost model for foodborne disease classification is evaluated using various evaluation criteria such as precision, recall, and fl-score. Each class has obtained its value of evaluation measures, as shown in Table 5.8.

Table 5.8 XGBOOST model evaluation on each class using precision, recall and f1_score

Classes	Train test Split	Precision	Recall	F1 Score
Amoebiasis	70/30	0.98	0.97	0.97
	80/20	0.97	0.97	0.97
Giardiasis	70/30	0.98	0.94	0.96
	80/20	0.98	0.94	0.96
Typhoid Fever	70/30	0.96	1.00	0.98
	80/20	0.96	0.99	0.98
Macro avg	70/30	0.97	0.97	0.97
	80/20	0.97	0.97	0.97
Weighted avg	70/30	0.97	0.97	0.97
	80/20	0.97	0.97	0.97

The XGBoost model on foodborne disease classification is evaluated for train and test accuracy to see how effectively the model was trained and classified foodborne diseases using unseen data. According to the findings of the XGBoost model experiment, the training accuracy is 99.45%, the test accuracy is 97.20% in 70/30 split and found 99.40% in train accuracy and 96.94% of test in 80/20 split, and also the overall performance of the model in terms of precision, recall and

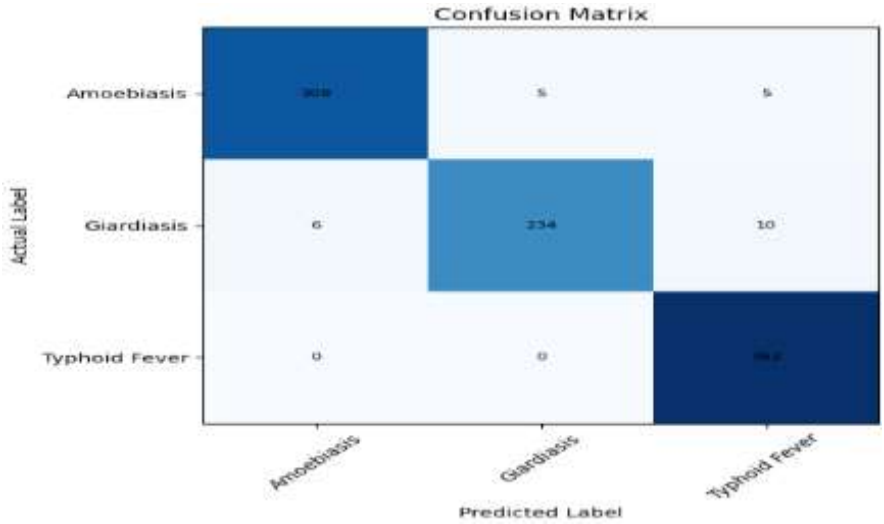
f1_score were conducted as shown in Table 5.9 below.

Table 5.9 Evaluation of XGBOOST model in accuracy, precision, recall and f1_score

Model name	Train test split	Precision	Recall	Fl_score	Train Accuracy	Test Accuracy
XGBoost	70/30	97.24%	97.20%	97.19%	99.45%	97.20%
	80/20	96.96%	96.94%	96.93%	99.40%	96.94%

In the confusion matrix, which is depicted in Table 5.7 below, the findings of the XGBoost model for the classification of foodborne diseases are enumerated. The XGBoost model classified 904 instances correctly out of 930 instances and incorrectly classified 26 instances. The interpretation of each classes are discussed as follow: The XGBoost model correctly classified 308 Amoebiasis instances of and incorrectly classified 10 instances (wrongly classified 5 instances as cases of Giardiasis and 5 instances as cases of Typhoid Fever), while the XGBoost model correctly classified all 334 instances of Giardiasis and wrongly classified 16 instances (wrongly classified 6 instances as Amoebiasis class and 10 instances as Typhoid Fever class), in Typhoid Fever class 362 instances are correctly classified and wrongly classified 0 instances. As per the result shown in Table 5.10, we can say Amoebiasis, Giardiasis and Typhoid Fever disease have shared symptoms in common.

Table 5.10 Confusion matrix for XGBOOST model



5.5.4 Stacking ensemble model

This study investigated the effectiveness of a stacking ensemble learning model for classifying foodborne diseases. We employed Decision Tree, Random Forest, and XGBoost as base learners. The hyperparameters for each base learner were optimized using grid search, and the chosen parameters and their corresponding values are presented in Table 5.1. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics in each class in both 70/30 and 80/20 train test splits as shown in Table 5.11.

Table 5.11 Stacking model evaluation on each class using precision, recall and f1_score

Classes	Train test Split	Precision	Recall	F1 Score
Amoebiasis	70/30	0.98	0.97	0.98
	80/20	0.99	0.97	0.98
Giardiasis	70/30	0.98	0.94	0.96
	80/20	0.98	0.96	0.97
Typhoid Fever	70/30	0.97	1.00	0.98
	80/20	0.97	1.00	0.99
Macro avg	70/30	0.98	0.97	0.97
	80/20	0.98	0.98	0.98
Weighted avg	70/30	0.98	0.98	0.98
	80/20	0.98	0.98	0.98

The Stacking model on foodborne disease classification is evaluated for train and test accuracy to see how effectively the model was trained and classified foodborne diseases using unseen data. According to the findings of the stacking model experiment, the training accuracy is 99.40%, the test accuracy is 97.53% in 70/30 split and found 99.35% in train accuracy and 98.06% of test in

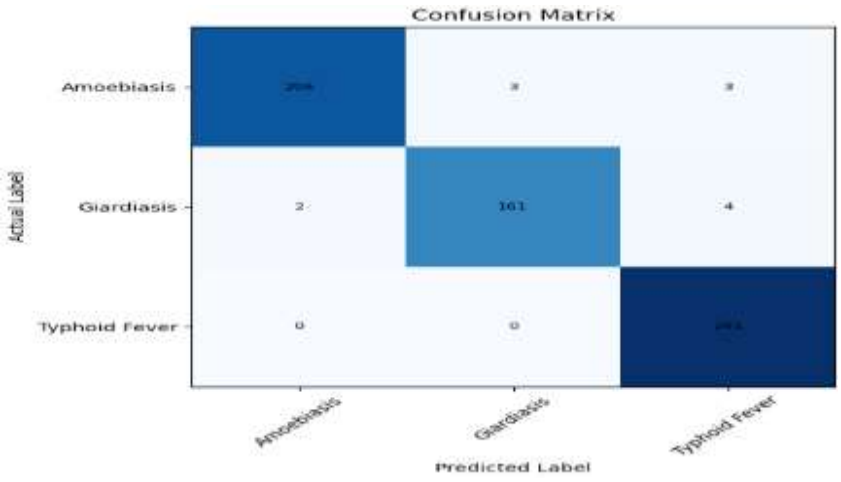
80/20 split, and also the overall performance of the model in terms of precision, recall and f1_score were conducted as shown in Table 5.12 below

Table 5.12 Evaluation of stacking model in accuracy, precision, recall and f1_score

Model name	Train test split	Precision	Recall	Fl_score	Train Accuracy	Test Accuracy
Stacking model	70/30	97.55%	97.53%	97.52%	99.40%	97.53%
	80/20	98.08%	98.06%	98.06%	99.35%	98.06%

In the confusion matrix, which is depicted in Table 5.13 below, the findings of the stacking model for the classification of foodborne diseases are enumerated. The stacking model classified 608 instances correctly out of 620 instances and incorrectly classified 12 instances. The interpretation of each classes are discussed as follow: The stacking model correctly classified 206 Amoebiasis instances of and incorrectly classified 6 instances (wrongly classified 3 instances as cases of Giardiasis and 3 instances as cases of Typhoid Fever), while the stacking model correctly classified all 161 instances of Giardiasis and wrongly classified 6 instances (wrongly classified 2 instances as Amoebiasis class and 4 instances as Typhoid Fever class), in Typhoid Fever class 241 instances are correctly classified and wrongly classified 0 instances. As per the result shown in Table 5.13, we can say Amoebiasis, Giardiasis and Typhoid Fever disease have shared symptoms in common.

Table 5.13 Confusion matrix for stacking model



5.6 Result Discussion and Comparative Analysis of all Models

In this section, the comparative analysis of all models is done and the outperforming and suitable model is selected for proposed model development based on the experiment result.

Based on the experiment conducted, each model obtained its accuracy value as follows: the decision tree model obtained 96.5%, random forest have 97.6%, XGBoost have 96.9% and Stacking ensemble learning method obtained 98.1% in a 80/20 split. According to the aforementioned result obtained on a 80/20 split and as depicted in Figure 5.2, Stacking ensemble learning model outperformed other models in terms of accuracy.

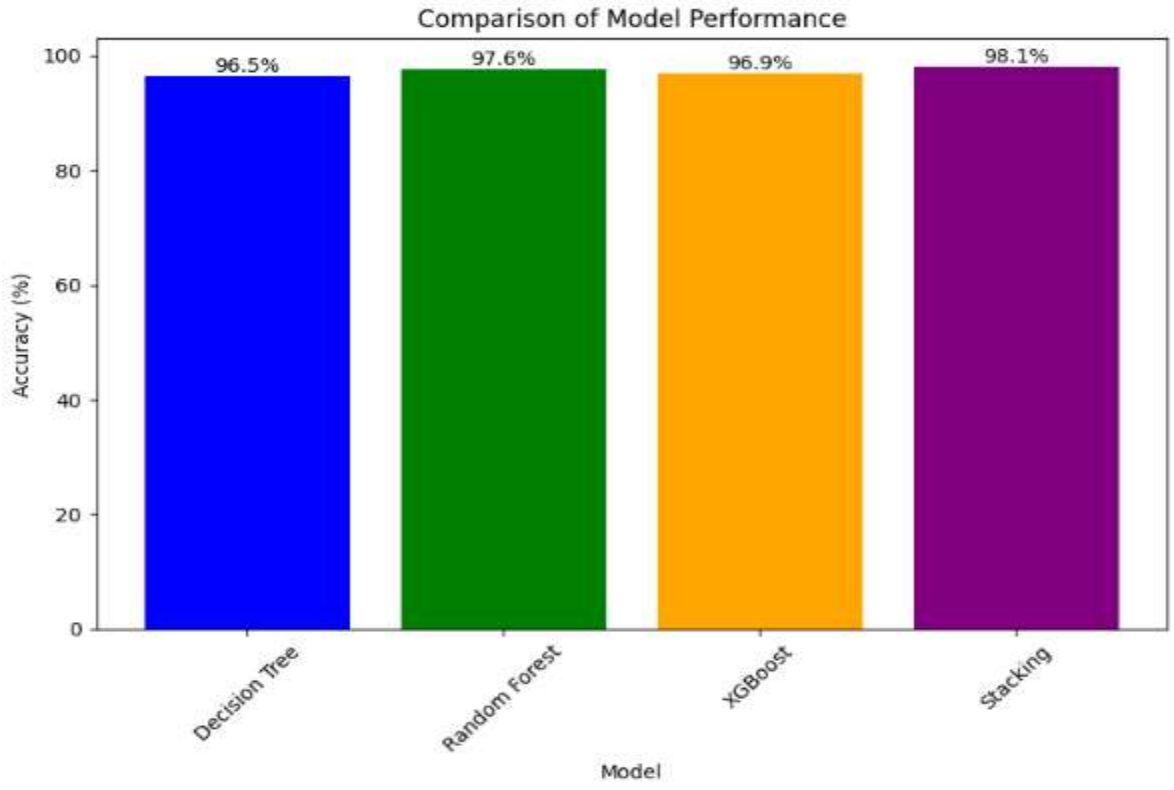


Figure 5.2 Comparison of all models in terms of accuracy

The precision, recall, and f1 score of all models are presented in Table 5.14 in order to evaluate how accurately the models trained and correctly classified foodborne diseases using unseen data in both the 80/20 and 70/30 split.

Table 5.14 Comparisons of all models in terms of precision, recall, and f1_score

Model	**Train test split**	**Precision**	**Recall**	**F1_score**
DT	70/30	0.9667	0.9667	0.9666
	80/20	0.9647	0.9645	0.9645
RF	70/30	0.9712	0.9710	0.9708
	80/20	0.9761	0.9758	0.9757
XGBOOST	70/30	0.9724	0.9720	0.9719
	80/20	0.9696	0.9694	0.9693
Stacking	70/30	0.9755	0.9753	0.9752
	80/20	**0.9808**	**0.9806**	**0.9806**

In model random forest and stacking , 80/20 splits outperform 70/30 splits in terms of precision, recall, and f1 score, and in this split Stacking ensemble learning model performs better among others in terms of all above-mentioned metrics as shown in Table 5.14 above and graphically shown in Figure 5.3.

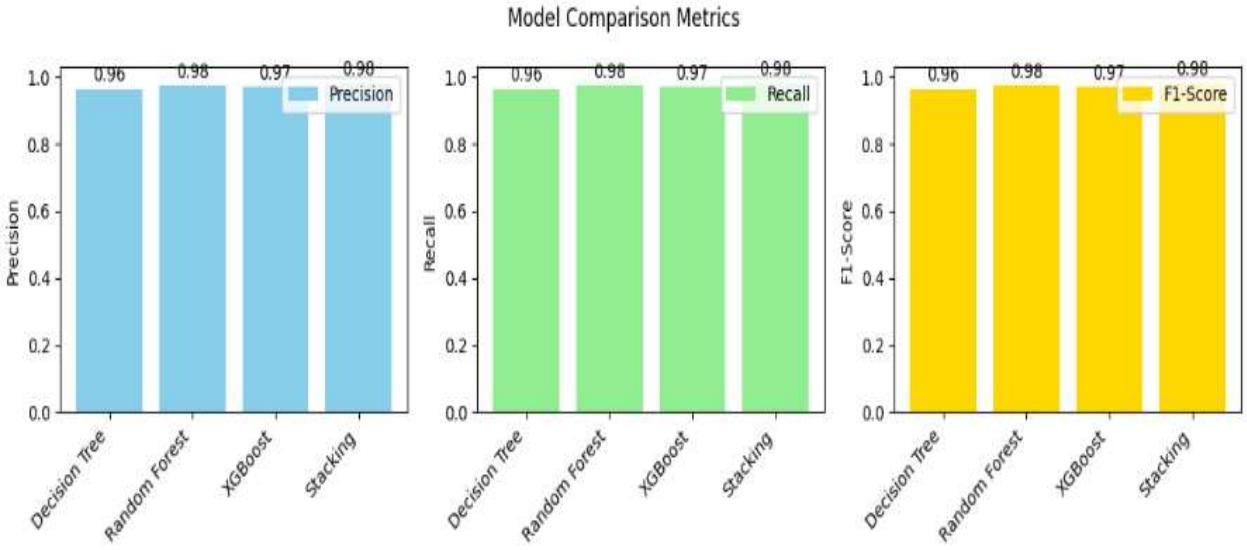


Figure 5.3 Comparison of all models in terms of precision, recall, and f1_score

The accuracy of each model is evaluated for train and testing to as certain how well it classified foodborne diseases using unseen data as shown in Table 5.15.

Table 5.15 Comparisons of all models in terms of train and test accuracy

Model	**Train test split**	**Train accuracy**	**Test accuracy**
DT	70/30	99.45%	96.67%
	80/20	99.44%	96.45%
RF	70/30	99.45%	97.10%
	80/20	99.44%	97.58%
XGBOOST	70/30	99.45%	97.20%
	80/20	99.40%	96.94%
Stacking	70/30	99.40%	97.53%
	80/20	**99.35%**	**98.06%**

The Stacking model outperforms others in terms of test accuracy as shown in Table 5.15 above in both the 70/30 split and 80/20 split. However, 80/20 outperforms the 70/30 split.

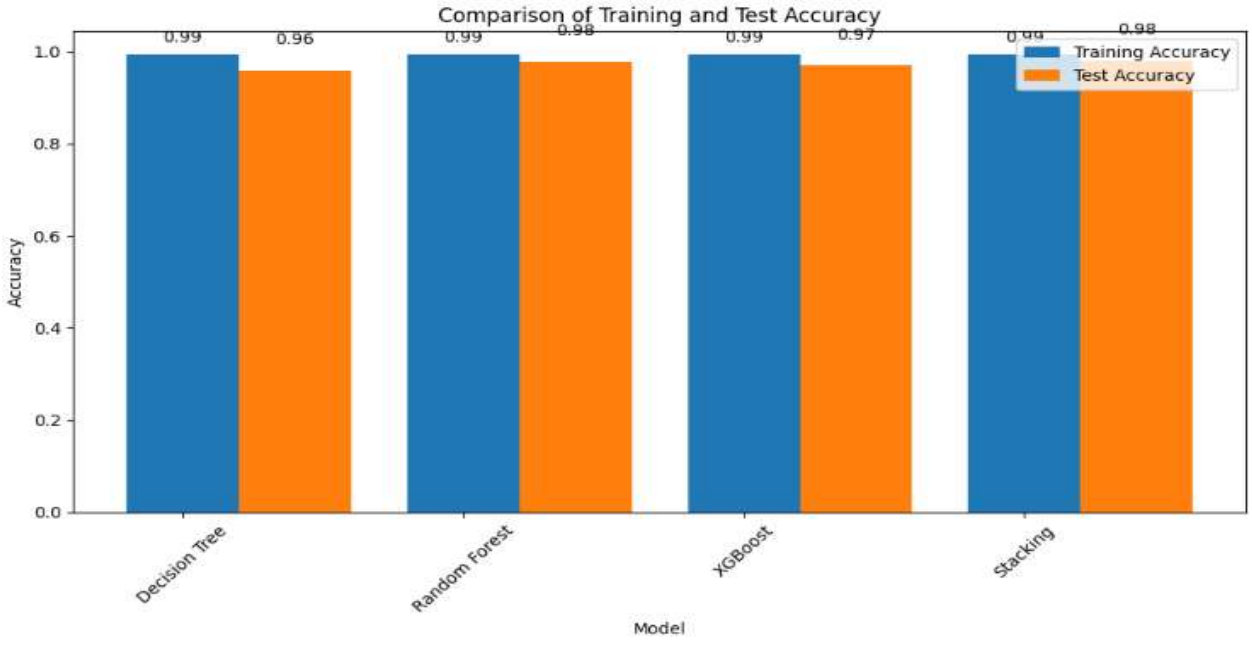


Figure 5.4 Comparison of all models in terms of train and test accuracy

Based on the above experimental results, the Stacking model outperformed other machine learning models such as Decision Tree, Random Forest, and XGBoost in terms of accuracy, precision, recall, and F1-score. Therefore, the researcher selected the Stacking ensemble learning model for the proposed model development due to its superior performance.

The reason behind the Stacking model outperforming others in all evaluation criteria is that its inherent ability to leverage the strengths of multiple base learners. In this research Stacking ensembles combined the classifications from selected models, which are decision trees, random forests, and XGBoost, to create a more robust and generalizable final model. This ensemble approach allowed the stacking model to capture different aspects of the data that individual base learners might miss. Furthermore, the hyperparameter tuning the researchers performed on each base learner using grid search likely contributed to the stacking model's success. By optimizing the hyperparameters of each individual model, the researcher ensured they functioned at their peak potential. These optimized base learners were then incorporated into the stacking framework, further enhancing the overall ensemble's performance. Finally, the researcher choice of logistic regression as the meta-learner in the stacking model might have also played a role. Logistic regression excels at learning linear relationships between features and the target variable. This characteristic might have been particularly effective in this case, allowing it to efficiently combine the diverse classifications from the tuned base learners, ultimately leading to the strong performance observed in above experiment.

Due to the aforementioned reasons and experiment conducted, the Stacking model outperforms other machine learning models like Decision Tree, Random forest, and XGBoost with an accuracy of 98.06%. Despite the fact that, the execution time of other machine learning algorithms takes less time than the stacking model, it obtained better performance in terms of accuracy, precision, recall, and f1_score. Therefore, we can say the Stacking model is a suitable model for foodborne disease classification and finally it is selected for proposed model development. As result, the question raised by a researcher under research question two is answered here.

In reviewing related literature, several studies have explored machine learning models for disease classification and prediction, particularly focusing on foodborne diseases like Typhoid Fever. Oguntimilehin et al [40] utilized Rough Set machine learning techniques to classify Typhoid Fever diseases, achieving a training set accuracy of 95% and a testing set accuracy of 96%. Similarly, Radhika et al[41] employed Decision Trees for disease classification based on symptoms, achieving accuracy, but their focus was broader and did not emphasize prevalent foodborne diseases. Wang et al [14] utilized various machine learning methods for classifying foodborne disease pathogens, with the gradient boost decision tree model achieving the highest accuracy. Furthermore, Leo et al[43] and Bhuiyan et al[45] explored machine learning algorithms for cholera and typhoid fever prediction, respectively, while Grampurohi et al[44] aimed to classify various diseases including Typhoid Fever. Additionally, Kumar et al[46] focused solely on predicting pathogens in foodborne diseases. However, none of these studies explicitly focused on combining multiple models through ensemble learning techniques like Random Forest, XGBoost, or Stacking to enhance classification accuracy. Therefore, the contribution of this research lies in its application of a Stacking ensemble learning model, leveraging the strengths of various base learners, hyperparameter tuning, and logistic regression as the meta-learner to achieve superior performance in terms of accuracy, precision, recall, and F1-score, thereby addressing the identified gap in the literature.

Chapter Six

Conclusion and Recommendation

6.1 Conclusion

As Ethiopia is a developing country, many individuals suffer from the effect of various foodborne diseases. Ethiopia has a high infection rate and most foodborne diseases are poorly controlled due to a lack of technology that can quickly classify foodborne diseases and aid medical personnel in making a more accurate diagnosis. There are some previous studies that tried to develop a classifier model for foodborne diseases. However, there is a lack of research done to classify the foodborne diseases that are widespread in Ethiopia. Thus, the researcher motivated to develop a classifier model that can easily classify foodborne diseases using machine learning algorithms.

To achieve the objective of the proposed study, the necessary data were collected from two Hospitals namely Gesuba primary hospital (GPH) and WSUCSH. The confirmed patient cases of foodborne diseases were collected based on purposive sampling and expert recommendations. The researcher used qualitative and quantitative research approaches while collecting the primary and secondary data from the hospitals. The qualitative data were taken from hospitals by interviewing the experts in hospital. 3100 secondary data were collected from both GPH and WSUCSH by recording the confirmed cases of foodborne diseases from 2012 to 2015 E.C. Experimental research design is applied in order to select a suitable machine learning model for the proposed study based on the experiments conducted. To prepare and analyze the collected data, different python libraries and modules such as Pandas, NumPy and Sklearn were used. After preprocessing the collected data, different state of art machine learning such as Decision Tree, Random Forest, XGBoost, and stacking ensemble learning model were applied to train a model and their performance was evaluated using various evaluation metrics such as confusion matrix, accuracy, precision, recall, and f1_score. Based on the evaluation made, among all trained models, the one which outperforms other models in terms of accuracy, precision, recall, and f1_score is selected for proposed model development. To train a model, 70/30 and 80/20 splits were applied and the 80/20 split results outperforms the 70/30 split in all aforementioned

evaluation criteria. In the 80/30 split, the Stacking ensemble model outperforms others with an accuracy of 98.1% followed by Random forest with an accuracy of 97.5%, XGBoost with 96.9%, and Decision Tree with 96.5%. Therefore the researcher selected the stacking ensemble model for proposed model development based on its good performance while training and testing. Finally, we can say the stacking ensemble model is suitable for foodborne disease classification based on the experiment conducted and its computational nature while training the data. Therefore, the researcher addressed the research questions raised and the research gap identified in the literature as follow:

In order to find an answer to research question 1(**What are the important attributes that are used to classify foodborne diseases**?), we collected both primary and secondary data on foodborne diseases. In order to know the importance of each feature, the secondary data was subjected to feature analysis using random forest feature importance. Based on feature importance analysis, some features obtained less score. So, instead of simply dropping the features which have less score, the researcher also made consultation with domain experts' to know the importance of those features in real scenario. To understand the relationship between each feature, the correlation coefficient matrix for each feature is also computed. As result, the researcher has able to identify the importance of each feature in a dataset. Therefore, the question raised under research question 1 has got an answer.

To answer research question 2(**Which machine learning model is suitable to classify foodborne diseases**?), the models were trained using state-of-art machine learning algorithms such as DT, RF, XGBoost, and stacking ensemble learning method, and their performance is evaluated using different evaluation metrics like accuracy, precision, recall, and f1_score. Then, the suitable model is selected based on evaluation and comparison made. Finally, the proposed model is developed. Based on the experiment conducted, the stacking ensemble learning model outperforms others and is selected for foodborne disease classification. In view of this fact, the question raised under research question 2 has got an answer. In addition to this, the researcher attempted to fill the research gap identified in the literatures by gathering secondary data on foodborne diseases (which are not included in the previous studies) that are prevalent in study area and developed a proposed model by using state of art machine learning model.

6.2 Recommendation

We planned to cover the foodborne diseases which are most prevalent in Ethiopia but the researcher couldn't cover all of them due to budget limits and lack of data on some foodborne diseases like Campylobacter Infections, rabies, anthrax, brucellosis, leptospirosis, echinococcosis, Shiga Toxin-Producing Escherichia coli (STEC). Thus, the researcher covered three foodborne diseases namely Amoebiasis, Giardiasis, and Typhoid Fever classification using machine learning and stacking ensemble learning model. In the future, the researcher would like to recommend for other researchers to classify all foodborne diseases prevalent in Ethiopia using Stacking ensemble machine learning algorithms.

6.3 Contribution

The proposed study contributed to the existing knowledge and practice by developing an appropriate machine learning model that can classify foodborne diseases. In addition to this, the study showed the direction for the future researcher to classify the foodborne diseases which are prevalent in Ethiopia. Moreover, we contributed to the future researcher by collecting and preparing the secondary data of foodborne diseases which is in manual format, and converted it to machine understandable format. Because, it takes a great effort to collect each confirmed cases of foodborne diseases from the patient history card and prepare it for model development. Finally, the proposed study delivered a classifier model for foodborne disease classification which can help healthcare professionals to make a better diagnosis.

Reference

[1] K. Hossain, D. Raheem, and S. Cormier, “Food Security: A Basic Need for Humans,” in *Food Security Governance in the Arctic-Barents Region*, Cham: Springer International Publishing, 2018, pp. 5–14. doi: 10.1007/978-3-319-75756-8_2.

[2] Mahon, C.R., “Foodborne illness: is the public at risk?,” *Clinical laboratory science : journal of the American Society for Medical Technology*, vol. 11 5, pp. 291–7, 1998.

[3] WHO Press, World Health Organization. *Foodborne disease outbreaks: Guidelines for investigation and control*. 20 Avenue Appia, 1211 Geneva 27, Switzerland (tel.: +41 22 791 3264; fax: +41 22 791 4857; e-mail: bookorders@who.int).

[4] L. Gazu., “Foodborne disease hazards and burden in Ethiopia: A systematic literature review, 1990–2019,” in *Frontiers in Sustainable Food Systems*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256938077

[5] A. Alanazi, “Using machine learning for healthcare challenges and opportunities,” *Informatics in Medicine Unlocked*, vol. 30, p. 100924, 2022, doi: https://doi.org/10.1016/j.imu.2022.100924.

[6] S. Mohammad, “Multi Disease Detection and Predictions Based On Machine Learning,” Sep. 2021.

[7] S. C. Teferi, “A Review on Food hygiene Knowledge, Practice and Food Safety in Ethiopia,” *Food Science and Quality Management*, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:229037211

[8] R.Agarwal, “The 5 Most Useful Techniques to Handle Imbalanced Datasets - KDnuggets,” 2020.

[9] J.Brownlee, “Stacking Ensemble Machine Learning With Python.” 2021.

[10] World Health Organization. “Foodborne diseases” [Online]. Available: https://www.who.int/health-topics/foodborne-diseases.

[11] Haile KT, Yabeyu AB, “Knowledge, Attitude and Practice of Pharmacy Professionals Against Dispensing Antibiotics Without Prescription in Ethiopia.,” *Integr Pharm Res Pract. 2022;11:167-176*, doi: https://doi.org/10.2147/IPRP.S383709.

[12] E. M. Koji, G. B. Gebretekle, and T. A. Tekle, “Practice of over-the-counter dispensary of antibiotics for childhood illnesses in Addis Ababa, Ethiopia: a simulated patient encounter

study," *Antimicrobial Resistance & Infection Control*, vol. 8, no. 1, p. 119, Jul. 2019, doi: 10.1186/s13756-019-0571-x.

[13] B. Ayele, Z. Mekonnen, T. Sisay Tessema, E. Adamu, E. Tsige, and G. Beyene, "Antimicrobial Susceptibility Patterns of Shigella Species among Children under Five Years of Age with Diarrhea in Selected Health Centers, Addis Ababa, Ethiopia.," *Can J Infect Dis Med Microbiol*, vol. 2023, p. 5379881, 2023, doi: 10.1155/2023/5379881.

[14] H. Wang, W. Cui, Y. Guo, Y. Du, and Y. Zhou, "Prediction of Foodborne Diseases Pathogens: A Machine Learning Approach (Preprint)," *JMIR Medical Informatics*, vol. 9, Oct. 2020, doi: 10.2196/24924.

[15] MOH-Ethiopia, "Health Sector Transformation Plan II HSTP II 2020/21-2024/25 (2013 EFY - 2017 EFY)," Feb. 2021.

[16] M.Clinic., "Shigella infection." Accessed: Jan. 30, 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/shigella/symptoms-causes/syc-20377529

[17] L. Vrbova, K. Johnson, Y. Whitfield, and D. Middleton, "A descriptive study of reportable gastrointestinal illnesses in Ontario, Canada, from 2007 to 2009," *BMC Public Health*, vol. 12, no. 1, p. 970, Nov. 2012, doi: 10.1186/1471-2458-12-970.

[18] M. Bush, "Shigellosis (Bacillary Dysentery)." Accessed: Jan. 30, 2024. [Online]. Available: https://www.merckmanuals.com/professional/infectious-diseases/gram-negative-bacilli/shigellosis

[19] World health organization (WHO), "*Estimating the burden of foodborne diseases: A practical handbook for countries*. " 2021.

[20] *World Health Organization*, "*EXECUTIVE SUMMARY*). " [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/200046/WHO_FOS_15.02_eng.pdf

[21] L. Gazu, "Foodborne disease hazards and burden in Ethiopia: A systematic literature review, 1990–2019," in *Frontiers in Sustainable Food Systems*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256938077

[22] L. Habte, E. Tadesse, G. Ferede, and A. Amsalu, "Typhoid fever: clinical presentation and associated factors in febrile patients visiting Shashemene Referral Hospital, southern Ethiopia," *BMC Research Notes*, vol. 11, no. 1, p. 605, Aug. 2018, doi: 10.1186/s13104-018-3713-y.

[23] Ethiopian Public Health Institute Center for Public Health Emergency Managemen, "Weekly Epidemiological Bulletin", May 2022, [Online]. Available: https://ephi.gov.et/wp-content/uploads/2021/12/EPHI_cPHEM_Weekly_Epidemiological_Bulletin_Epi-Week-17_2022-1.pdf

[24] G. B. Roro et al., "Prevalence and Associated Risk Factors of Entamoeba Histolytica Infection Among Schoolchildren From Three Primary Schools In Arsi Town, West Zone, Ethiopia," vol. PREPRINT (Version 1) available at Research Square, Jan. 2022, doi: https://doi.org/10.21203/rs.3.rs-1202313/v1.

[25] T. L. Switaj, K. J. Winter, and S. R. Christensen, "Diagnosis and Management of Foodborne Illness.," *Am Fam Physician*, vol. 92, no. 5, pp. 358–365, Sep. 2015.

[26] S. Angra and S. Ahuja, "Machine learning and its applications: A review," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 2017, pp. 57–60. doi: 10.1109/ICBDACI.2017.8070809.

[27] GeeksforGeeks, "Machine Learning." Accessed: Dec. 28, 2024. [Online]. Available: https://www.GeeksforGeeks.com/machine-learning

[28] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 945–949. doi: 10.1109/ICCONS.2018.8663155.

[29] GeeksforGeeks, "Supervised Machine Learning." [Online]. Available: https://www.geeksforgeeks.org/supervised-machine-learning/

[30] M.W. Berry, A. Mohamed, and B.W. Yap, "Supervised and unsupervised learning for data science," *Springer*, 2019.

[31] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A Simple Semi-Supervised Learning Framework for Object Detection," *CoRR*, vol. abs/2005.04757, 2020, [Online]. Available: https://arxiv.org/abs/2005.04757

[32] C. Szepesvári, "Algorithms for reinforcement learning," *Springer Nature*, 2022.

[33] C. A. UI Hassan, M. S. Khan, and M. A. Shah, "Comparison of Machine Learning Algorithms in Data classification," in *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748995.

[34] Towards Data Science, “Decision Trees for Classification—Complete Example,” *Towards Data Science*, Jan. 2023, Accessed: Apr. 17, 2024. [Online]. Available: https://towardsdatascience.com/decision-trees-for-classification-complete-example-d0bc17fcf1c2

[35] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, “A comparison of random forest variable selection methods for classification prediction modeling,” *Expert Systems with Applications*, vol. 134, pp. 93–101, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.028.

[36] Binay Gupta, “Random Forest Algorithm in Machine Learning.” Accessed: Feb. 29, 2024. [Online]. Available: https://www.scaler.com/topics/machine-learning/random-forest-algorithm/

[37] Bex T, “Beginner’s Guide to XGBoost for Classification Problems” Accessed: Feb. 29, 2024. [Online]. Available: https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

[38] David Martins, “XGBoost: A Complete Guide to Fine-Tune and Optimize your Model | by David Martins | Towards Data Science.” Accessed: Feb. 29, 2024. [Online]. Available: https://towardsdatascience.com/xgboost- fine-tune-and-optimize-your-model-23d996fab663

[39] Yash Khandelwal, “Ensemble Stacking for Machine Learning and Deep Learning,” Aug. 2021, Accessed: Mar. 26, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/

[40] Oguntimilehin, A. and Adetunmbi, A.O. and Abiola, O.B., “A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever,” 2013, [Online]. Available: http://eprints.abuad.edu.ng/id/eprint/73

[41] S. Radhika et al., “Symptoms based disease prediction using decision tree and electronic health record analysis,” *European Journal of Molecular \& Clinical Medicine*, 2020.

[42] T. Kavzoglu, A. Teke, “Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost),” *Arabian Journal for Science and Engineering*, vol. 47, no. 6, pp. 7367–7385, Jun. 2022, doi: 10.1007/s13369-022-06560-8.

[43] J. Leo, E. Luhanga, K. Michael, "Machine Learning Model for Imbalanced Cholera Dataset in Tanzania," *The Scientific World Journal,* vol. 2019, p. 12, 2019, doi: https://doi.org/10.1155/2019/9397578.

[44] S. Grampurohit, C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," p. pp 1-7, Jun. 2020.

[45] Md. A. Bhuiyan, S. S. Rad, F. T. Johora, A. Islam, M. I. Hossain, and A. A. Khan, "Prediction Of Typhoid Using Machine Learning and ANN Prior To Clinical Test," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 2023, pp. 1–7. doi: 10.1109/ICCCI56745.2023.10128226.

[46] Y. Kumar, I. Kaur, and S. Mishra, "Foodborne Disease Symptoms, Diagnostics, and Predictions Using Artificial Intelligence-Based Learning Approaches: A Systematic Review," *Archives of Computational Methods in Engineering*, Aug. 2023, doi: 10.1007/s11831-023-09991-0.

[47] Zach Bobbitt, "Label Encoding vs. One Hot Encoding: What's the Difference?," Aug. 2022, Accessed: Apr. 03, 2024. [Online]. Available: https://www.statology.org/label-encoding-vs-one-hot-encoding/

[48] Statistics Globe, "Mode Imputation (How to Impute Categorical Variables Using R)." [Online]. Available: https://statisticsglobe.com/mode-imputation/

[49] T. A. Assegie, S. J. Sushma, B. G. Bhavya, and S. Padmashree, "Correlation Analysis for Determining Effective Data in Machine Learning: Detection of Heart Failure," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–5, 2021, doi: doi: 10.1007/s42979-021-00617-5.

[50] Akash Dubey, "Feature Selection Using Random forest | by Akash Dubey | Towards Data Science." Accessed: Mar. 19, 2024. [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest- 26d7b747597f

[51] A. Aylin Tokuç, "Splitting a Dataset into Train and Test Sets | Baeldung on Computer Science." Accessed: Mar. 19, 2024. [Online]. Available: https://www.baeldung.com/cs/train-test-datasets-ratio

[52] Rahul Kumar Sevakula, Wan-Tai M. Au-Yeung, Jagmeet P. Singh, E. Kevin Heist, Eric M. Isselbacher and Antonis A. Armoundas, "State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System," *Journal of*

the American Heart Association., Feb. 2020, [Online]. Available: https://doi.org/10.1161/JAHA.119.013924

[53] Mukesh Mithrakumar, "How to tune a Decision Tree?," *Towards Data Science*, Nov. 2019, [Online]. Available: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

[54] GeeksforGeeks, "Random Forest Hyperparameter Tuning in Python," www.geeksforgeeks.org. Accessed: Mar. 21, 2024. [Online]. Available: https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/

[55] Rahul Shah, "Tune Hyperparameters with GridSearchCV," Dec. 2023, Accessed: Mar. 21, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/

[56] Eric Luellen, "Mastering XGBoost. Hyper-parameter Tuning & Optimization | by Eric Luellen | Towards Data Science", Accessed: Mar. 21, 2024. [Online]. Available: https://towardsdatascience.com/mastering-xgboost- 2eb6bce6bc76

[57] J. Chen, A. Zeb, Y. A. Nanehkaran, and D. Zhang, "Stacking ensemble model of deep learning for plant disease recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12359–12372, Sep. 2023, doi: 10.1007/s12652-022-04334-6.

[58] Developers , "Classification: Accuracy," developers.google.com. Accessed: Mar. 22, 2024. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy

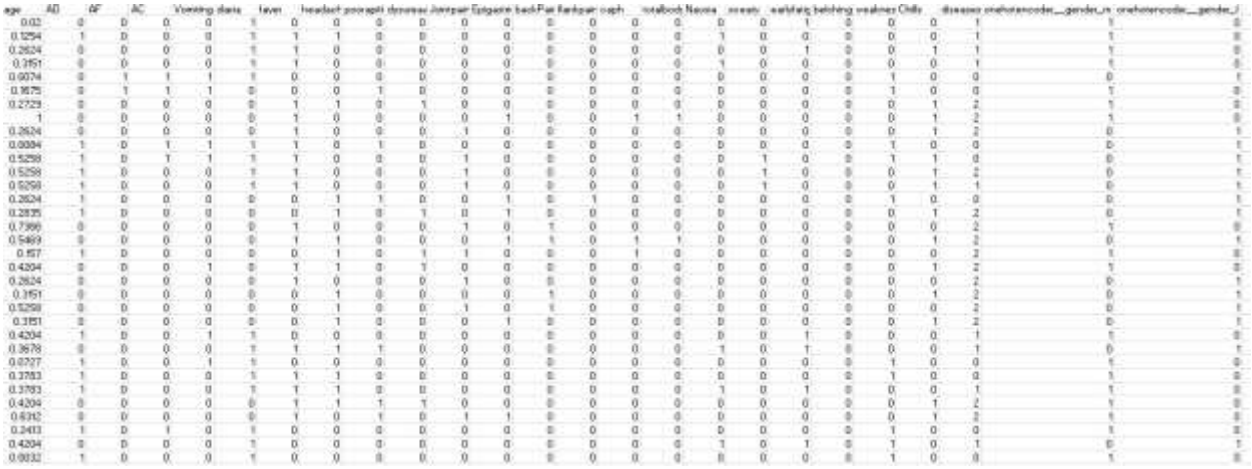
[59] Rosaria Silipo, "Machine Learning Algorithms and The Art of Hyperparameter Selection", Accessed: Apr. 09, 2024. [Online]. Available: https://towardsdatascience.com/machine-learning-algorithms-and-the-art-of- hyperparameter-selection-279d3b04c281

APPENDICES

Appendix A: Sample Dataset of Foodborne Disease before Preprocessing

age	gender	AD	AF	AC	Vomiting	diaria	faver	headachi	poorapiti	dysurea	Jointpair	Epigastri	backPain	flankpair	caph	totalbod	Naosia	sweaty	earlyfatig	belching	weaknes	Chills	diseases
2	m	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gardiasis
12	m	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gardiasis
25	m	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Gardiasis
30	m	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gardiasis
0.8	f	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Amebiasis
16	m	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	Amebiasis
26	m	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Typhoid
95	m	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	Typhoid
25	f	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Typhoid
0.9	f	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Amebiasis
50	f	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	Amebiasis
50	f	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	Typhoid
50	f	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	Gardiasis
25	f	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	Amebiasis
27	f	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	Typhoid
70	m	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	Typhoid
52	f	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	Typhoid
15	m	1	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	Typhoid

Appendix B: Sample Dataset after Preprocessing



Appendix C: Sample Python Libraries used for Proposed Model Development

```
import numpy as np

import pandas as pd

import seaborn as sns

import plotly.express as px

from matplotlib import pyplot as plt

from sklearn.model_selection import cross_val_score

from collections import Counter

from sklearn.preprocessing import OneHotEncoder

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import MinMaxScaler

from sklearn.preprocessing import FunctionTransformer

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.metrics import confusion_matrix, recall_score, precision_score, f1_score,
accuracy_score, roc_auc_score

from sklearn.metrics import classification_report

from sklearn.metrics import roc_auc_score, RocCurveDisplay

from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import StackingClassifier

from sklearn.ensemble import RandomForestClassifier
```

```
from xgboost import XGBClassifier

from sklearn.linear_model import LogisticRegression
```

Appendix D: Data Preprocessing

```
# Checking the number of missing values in each column

data.isnull().sum()
```

#Handling missing values

```
# Columns to impute with mode

impute_cols = ['AF', 'diaria', 'poorapitite', 'earlyfatigue', 'belching', 'weakness', 'Chills']

# Function to impute missing values with the mode (most frequent value)

def impute_mode(df, cols):

    for col in cols:

        df[col].fillna(df[col].mode()[0], inplace=True)

    return df

# Impute missing values in the copy of the DataFrame

df_imputed = impute_mode(data.copy(), impute_cols)

# Print the imputed DataFrame

print(df_imputed)
```

#**min max Normalization**

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

df[['age']] = scaler.fit_transform(df[['age']])
```

Appendix E: Experiments for hyper parameter selection using Grid search

```
from sklearn.model_selection import GridSearchCV

from sklearn.ensemble import RandomForestClassifier

param_grid = {

    'n_estimators': [100,150, 200, 300],

    'max_depth': [5, 10, 15,20,24,None],

    'min_samples_split': [2, 5, 10],

    'min_samples_leaf': [1, 2, 4],

    'max_features': ['sqrt', 'log2', None],

    'criterion': ['gini', 'entropy']

}

# Create a Random Forest Classifier instance (adjust random_state if needed)

model = RandomForestClassifier(random_state=0)

# Set up GridSearchCV with appropriate scoring metric

grid_search = GridSearchCV(estimator=model, param_grid=param_grid, scoring='accuracy',
cv=5, n_jobs=-1)

# Fit the grid search to the training data

grid_search.fit(df_trainX, df_trainY)

# Print the best parameters and score

print("Best Parameters:", grid_search.best_params_)

print("Best Cross-Validation Score:", grid_search.best_score_)

# Access the best model from GridSearchCV
```

```
best_model = grid_search.best_estimator_
```

Appendix F: Sample Code Selected Model

```
estimators = []

estimators.append(('DecisionTree', DecisionTreeClassifier(criterion= 'gini', max_depth= 23, max_features= None, min_samples_leaf= 1, min_samples_split= 2,random_state=101) ))

estimators.append(('RandomForest', RandomForestClassifier(criterion='gini', max_depth=15, max_features= 'sqrt', min_samples_leaf= 1, min_samples_split=2, n_estimators= 200,random_state=101)))

estimators.append(('XGB', XGBClassifier(objective='multi:softprob',colsample_bytree= 0.5,gamma=0.0, learning_rate=0.1, max_depth=15, min_child_weight=1,random_state=101)))

meta_learner=LogisticRegression(multi_class='multinomial')

from sklearn.ensemble import StackingClassifier

SC = StackingClassifier(estimators=estimators,final_estimator=meta_learner)

SC.fit(df_trainX, df_trainY)

y_pred = SC.predict(df_testX)

print(f"\nStacking classifier training Accuracy: {SC.score(df_trainX, df_trainY):0.2f}")

print(f"Stacking classifier test Accuracy: {SC.score(df_testX, df_testY):0.2f}")
```

Appendix G: Ethical Clearance from Hawassa University IRB.

ሀዋሳ ዩኒቨርሲቲ
ህክምናና ጤና ሳይንስ ኮሌጅ
የምርምር ስነ-ምግባር ጉምጋሜ ቦርድ

HAWASSA UNIVERSITY
COLLEGE OF MEDICINE AND HEALTH SCIENCES
Institutional Review Board

Ref. No: IRB/105/16
Date: 14/03/2024

Name of Researcher(s): **Wuletaw Iyasu, Digif Teka (PhD)**

Topic of Proposal: ***Developing machine learning model for foodborne diseases classification***

Dear researcher(s),
The Institutional Review Board (IRB) at the College of Medicine and Health Sciences of Hawassa University has reviewed the aforementioned research protocol with special emphasis on the following points:

1. Are all principles considered?
   - 1.1. Respect for persons: Yes ☑ No ☐
   - 1.2. Beneficence: Yes ☑ No ☐
   - 1.3. Justice: Yes ☑ No ☐
2. Are the objectives of the study ethically achievable? Yes ☑ No ☐
3. Are the proposed research methods ethically sound? Yes ☑ No ☐

Based on the aforementioned ethical assessment, the IRB has:

- **A.** Approved the proposal for implementation ☑ Approval period -**14 MAR. 2024 to 13 MAR. 2025**
- **B.** Conditionally Approved ☐ Element Approved: **Protocol Version No. 1**
- **C.** Not Approved ☐ Follow up report expected in **6 months**

Obligation of the PI:

1. Should comply with the standard international and national scientific and ethical guidelines
2. All amendment and changes made in protocol and consent form needs IRB approval
3. The PI should report SAE within 3 days of the event
4. End of the study, including the manuscript should be reported to the IRB

Yours faithfully,

[illegible]
[illegible]
[illegible]

[illegible] Fax: +046 2208755 [illegible]

Appendix H: Sample Interview Questions for Domain Expert

1. What are the most prevalent foodborne diseases in the study area?
2. What are the factors that contribute to the high burden of foodborne diseases?
3. Which foodborne diseases have shared symptoms in common?
4. How to identify the foodborne diseases which have shared symptoms in common?
5. From which area the patients are come for foodborne diseases treatment to GPH and WSUCSH?
6. In which age group and gender, the foodborne disease is prevalent in the study are?